THE UNIVERSITY OF ALBERTA

AN EMPIRICAL INVESTIGATION OF THE SAMPLING DISTRIBUTION OF THE RELIABILITY COEFFICIENT ESTIMATES BASED ON ALPHA AND KR20 VIA COMPUTER SIMULATION UNDER VARIOUS MODELS AND ASSUMPTIONS

by

© KYUNG SUN BAY

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND

RESEARCH IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

FALL, 1971

ABSTRACT

The exact sampling distribution of reliability estimates of a composite test is known only for the case when the test scores of the parts can be expressed in a linear model and satisfy all the assumptions of the two way mixed model ANOVA with one observation per cell including normality of true and error scores. For more general cases, the sampling distributions have been in general unknown or ignored by the psychometricians.

This study examined the more liberal concepts of test theory and reliability in terms of the underlying models and assumptions, and investigated the sampling distribution of reliability estimates by performing a number of computer simulated sampling experiments under various models and distributional assumptions for true and error scores. The models employed were a mixed model ANOVA, essentially τ equivalent measurements, congeneric and multi-factor true score models for continuous cases, and the normal ogive model for binary item cases. For the distribution of true or latent and error scores, uniform, normal and exponential distributions were used.

The most general model was found to be a multi-factor true score model and all others could be shown to be special cases of this model. The most important factors influencing the sampling distribution are found to be uni-factorness and normality of true scores for continuous cases, and homogeneity of item difficulty parameters for binary cases. The distributions of error scores were found to be unimportant for both cases.

To determine the robustness conditions of the traditional F-test, the empirical distributions obtained by the sampling experiments were compared with those theoretical distributions obtained under the ANOVA and normal theory model. A number of conditions for robustness are given.

A new formula for the standard error of reliability estimates is introduced by analytical means and the validity of the formula was examined through computer simulated experiments. The new formula was found to be superior to traditional formulas based on normal theory when the normality of true score is not valid. Though the formula is derived under the ANOVA model, it was also found to be better than the traditional formulas under more general models.

Implications of these findings to test theory and applications are discussed and some numerical examples are given to show how the findings and the computer programs developed might be applied in practical situations.

ACKNOWLEDGEMENTS

The author wishes to express his deep and sincere appreciation to Dr. S.M. Hunka, under whose advice and guidance this study has been completed. Thanks are also expressed to other committee members: Drs. T.O. Maguire, E.S. West, and A.R. Hakstian.

Appreciation is also due to Dr. D.P. Flathman, Mr. J. Hanson of the Alberta Human Resources Research Council, Mr. D. Precht and other colleagues of the Division of Educational Research Services of the University of Alberta for their assistance and encouragement.

The author wishes to acknowledge with gratitude the financial assistance received from the University of Alberta through the Division of Educational Research Services, Faculty of Education.

Thanks are also extended to the typist, Miss J. Talpash, for her willingness to comply with the complexities of notations and to meet the necessary deadlines.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1.0 The General Problem

The estimation and interpretation of reliability has been a central issue for psychometric theorists and test authors as well as the users of educational and psychological tests. The reliability coefficient of a test, a population parameter, is defined as the ratio of the true score variance due to individual differences of the subjects, to the total test score variance in a population for which the test is developed.

A number of formulas for measuring reliability have been derived by many theorists since the initial formulation by Spearman (1910) of his theory of true and error scores. Most of the formulas express reliability as a function of the moments of the part scores and the total test scores under assumptions of parallel or equivalent measurements among the part tests, or, as a correlation coefficient between the observed test scores and a second set of scores on a real or hypothetical variable. In most cases, the formulas involve only point estimation of the reliability, and have been obtained by substituting the sample moments of the part-scores and the total scores into formulas which are valid in the population. Statistical properties of such estimates are in general unknown or ignored.

Investigation of the distribution of reliability estimates requires a mathematical model and a number of assumptions. The validity of the estimates of reliability largely depends on the validity

of the model and underlying assumptions. Even for rather rigorous models and assumptions the sampling distributions are unknown except in some special cases.

For a valid statistical inference about a population parameter the sampling distribution of the parameter estimate must be known, and the reliability cannot be an exception. For example, if a standardized test has been administered to a sample of subjects, it is sometimes necessary to compare the sample reliability estimate with the reliability claimed by the test authors, i.e., it is desirable to know whether the difference between the two values can be attributed to sampling fluctuations, or, whether there is a significant difference due to population difference. If a test is administered to two independent samples of subjects, a comparison of the two estimates of reliability may be necessary to determine the underlying cause of any observed difference.

The standard error of the reliability estimate is another useful measure of the precision of the estimates, but without any knowledge of the sampling distribution of the estimates, confidence intervals for the population reliability are impossible to calculate.

Most of the available formulas for reliability estimate depend on the estimation of variance components, using various, explicit or implicit, parallel or equivalent test form assumptions among the part test scores. Even though the estimates of the variance components thus obtained are usually unbiased, the estimates of the reliability are, in most cases, biased, and the statistical properties are unknown.

Since the early years of test theory, it has been

recognized by theorists and test users that the calculated reliability is, in fact, nothing more than an estimate of the true or population reliability, and therefore subject to sampling fluctuations. Even with this recognition, little work has been done to investigate the distribution of such estimates.

This study will investigate properties of the sampling distribution of reliability estimates based on Alpha or KR20 formulas using computer simulation techniques, and will employ various models and distributional assumptions for true or latent, and error scores described in the following two chapters.

1.2.0 Review of Related Studies

1.2.1 Concepts of Reliability

Even during the initial developments of test theory, psychologists showed interest in the formula for the standard error of reliability estimates as an indicator of the precision of such estimates. During this period most psychologists interpreted reliability as a correlation coefficient between classically defined parallel measures. Using this definition, attempts were made to apply the well known sampling distribution of correlation coefficient with the assumption of a bivariate normal distribution. However, unlike the usual inference about correlation coefficients, in most cases, the population reliability is considered to be close to unity rather than zero, and hence its distribution has extreme negative skewness making the usual normal approximation of little use (Jackson and Ferguson, 1941, p. 12).

When the split half method was introduced, the reliability

estimate was seen to depend on the way the test was split. As a result, the reliability estimate based on Alpha or KR20 was considered to be superior to the split half estimate since the former gave a unique estimate.

Cronbach (1951) has shown that Alpha or KR20 is an average of all possible split half reliabilities in the population. He thoroughly investigated the coefficient Alpha from the point of view of factorial structure. The Alpha coefficient was interpreted as the proportion of the test variance due to all common factors among the part scores, and as an index of consistency, an estimate of first factor concentration. He also showed that Alpha is a lower bound of the test reliability, but did not explicitly discuss the sampling aspect of the Alpha estimate.

The concept of test reliability has been under continuous change: the classical concept based on parallel tests has been modified and the assumptions relaxed. Burt (1955) and Tryon (1957) initiated a new concept of domain sampling, and the reliability as an index of generalizability has been advocated by Rajaratnam (1960), Cronbach, Rajaratnam and Gleser (1963), and Rajaratnam, Cronbach and Gleser (1965). Their conceptual framework relied heavily on ANOVA models, and initiated a process of liberalization of reliability theory from the rather restrictive classical orthodoxy of test parallelism. However their efforts concentrated on the problem of point estimation, and little attention was paid to sampling aspect of the estimates.

Lord and Novick (1968, p. 50) defined the concept of 'essentially τ equivalent measurements', and Novick and Lewis (1967)

have shown that the coefficient Alpha is identical to the reliability coefficient if and only if a test consists of essentially τ equivalent parts. If this condition is not satisfied Alpha is a lower bound for the reliability, confirming previous studies of Guttman (1945, 1953), Cronbach (1951), and others. To evoke the principle of essentially τ equivalent parts, it has been argued that only true score variances need be identical, i.e., homogeneity of true score variances, and not identical error score variances nor identical true scores among the part tests. The assumptions of classical parallel tests are, therefore, relaxed substantially. Jöreskog (1968, 1970, 1971) defines the concept of congeneric test scores which measure the same trait except for errors, relaxing the essentially τ equivalent measurement conditions further. Under this model any pair of such tests have linearly related true scores. The sampling distributions of the reliability estimates under these models are not yet generally known.

As an alternative to conventional uni-factor true score models, a multi-factor true score model has been advocated by LaForge (1965) using the multiple factor analysis model. An estimate of the squared multiple correlation of a part score with the scores of remaining parts, which is one estimate of the test communality in factor analysis, is proposed as an estimate of the reliability of the part score.

For certain kinds of tests, especially in the field of achievement tests, this approach seems more reasonable than the conventional uni-factor approach, but the old controversial problem of determining the number of factors in a factor analysis must still be resolved. However, the multi-factor model provides a general model

for computer simulation purpose, as will be seen in the next chapter, since the ANOVA and other models may be considered as special cases of the multi-factor model.

1.2.2 Sampling Theories of Reliability Estimates

Lord (1955) defined three kinds of sampling arising in test theory; sampling of subjects (Type 1), part tests or items (Type 2), and a simultaneous combination of the two (Type 12). Lord also discussed the sampling distribution of KR20 under Type 2 sampling without the presentation of the standard error of the KR20 estimates in terms of the population parameters.

A statistical sampling theory of the reliability estimates has been made possible through the application of ANOVA techniques to test theory. Hoyt (1941), Jackson and Ferguson (1941), Ebel (1951), Burt (1955), Cronbach, Rajaratnam and Gleser (1963), Feldt (1965, 1969), Maguire and Hazlett (1969) and many others investigated the reliability estimate under some form of ANOVA models. However, most of their discussion was limited to point estimation and little attention has been paid to the sampling fluctuation of the estimates or interval estimates.

Since the ANOVA models usually provide unbiased, consistent estimates of the variance components by some linear combination of various mean squares, the reliability estimates thus obtained are in general consistent estimators. But, in most cases, they are biased and do not have the desirable minimum variance property.

Although Jackson and Ferguson (1941, p. 40) related the F-statistic to the so called 'sensitivity of a test', or the square

root of the commonly referred signal-noise ratio, it was Ebel (1951) who first explicitly linked the sampling distribution of the reliability estimate itself to an F-statistic. He applied the concept of 'intra-class' correlation coefficient to a rating data set, and by employing the well known F-distribution, has shown a way to obtain confidence intervals of the population intra-class correlation which he interpreted as the reliability of a judge. However, the assumptions underlying the ANOVA model were not explicitly specified.

Kristof (1963) presented a rather complete sampling theory of reliability estimates within the context of the assumptions of classical reliability theory with the exception that the means of the part tests were allowed to be different, i.e., the part test scores are 'essentially' parallel measurements. Under Type 1 sampling and the assumption of a multi-normal distribution of the part test scores, a maximum likelihood estimator of the common correlation among the parts was obtained, i.e., the intra-class correlation coefficient. It was shown to be biased. A bias-free formula was introduced and the sampling distribution of the estimates based on this formula is shown to be related to the F-statistic. A method of statistical inference about the intra-class correlation, which was interpreted as the reliability of a part test, was suggested. Kristof's results are in close agreement with those obtainable under an ANOVA model. He has also showed that the estimate of the Alpha coefficient, in terms of second moment sample statistics, is the same as the maximum likelihood estimator of the reliability when a test has been divided into essentially parallel parts, with an assumption that the parts have a multi-normal distribution.

Kristof (1964) also investigated the distribution of reliability estimates for the first time without relying on the classical equal variance assumptions among the part test scores. A working formula for testing the significance of the difference between the two reliability estimates was derived under the assumption that each part has been administered to the same sample of subjects and that each part test could be split into parallel halves. He also investigated (1969, 1970) the sampling distribution of reliability estimates under the multi-normal assumptions when a test has been split into two parts not necessary parallel in the classical sense. A likelihood ratio test of the point hypothesis concerning the population value of Alpha was derived. This method was then used to yield confidence intervals for the parameter for any chosen level of confidence.

For the case where the parts of a test are simply binary items, the sampling distribution of KR20 estimates is more complicated than the case of continuous part scores. Most theorists have assumed an intermediate hypothetical variable between the item response and underlying true or latent trait score and linked the two variables with the help of the intermediate variable and a mathematical model. Lord (1952), and Lord and Novick (1968) used a normal ogive model, while Birnbaum (1967, 1968) proposed logistic, Poisson and other mathematical models. Although the latent trait approach provides means for investigation of the relationships among the item parameters and the test scores, nothing analytical has yet been done for the distributional theory of reliability estimates or its application even with the restrictive mathematical models and assumptions.

Aoyama (1957) has given explicit formulas, in terms of population parameters without any distributional assumptions, for the expected value and variance of KR20 estimates for Type 1 and Type 2 sampling situations. These results clearly indicate that the estimates are biased. However the formulas involve some approximations and calculation of higher order moments, and are too complex for any practical use.

Since the exact sampling distribution of KR20 estimates is not obtainable by analytical means, some researchers have attempted to approximate it by an ANOVA model. Feldt (1965) has investigated the applicability of the ANOVA model. He pointed out that imposition of a one-zero scoring scheme violates such assumptions of the ANOVA model as continuity of the scores, homogeneity of error variances and independence of true and error scores. He compared the results obtained under the ANOVA model with an empirical distribution based on real data reported by Baker (1962), and claimed the model robust when the assumptions are violated. Feldt referred to the model as a two way random effects model, but actually it was a mixed model as will be seen in the following chapter. Further applications were made of the method by deriving a scheme for testing the equality of two KR20 coefficients based on two independent samples using an approximate distribution of the product of two independent F-statistics (Feldt, 1969). Cleary and Linn (1968) adopted the same method as Feldt and gave an explicit formula for the standard error of KR20 estimates. However, their results are heavily dependent on normality assumptions which are not satisfied for KR20 cases.

Except for the case of approximation by employing unrealistic

assumptions of the ANOVA model for essentially parallel tests, most of the attempts to obtain the sampling distribution of KR20 estimates have either failed or resulted in unuseable formulas such as given by Aoyama (1957).

1.2.3 Empirical Approaches

The use of empirical approach to solve a statistical problem is as old as statistics itself. For example "Student" (1908) derived the analytic expression for the t-statistic and also established the validity of his argument by a sampling experiment. In education and psychology, a number of empirical investigations have been performed, with or without the help of a computer, to ascertain the robustness of the F-test when certain assumptions underlying an ANOVA model are not satisfied. Norton (1950), Boneau (1960), Hsu and Feldt (1969), and Bay (1970) are some examples of such investigations.

In reliability theory, Baker (1962) investigated a sampling distribution of KR20 estimates under Type 1 sampling constraints by actually performing experiments using real test results. Payne and Anderson (1968) tabulated the sampling distribution of KR20 estimates based on computer simulation. However, their experiments were limited to the cases of equal item difficulty parameters and inter-item correlations, i.e., phi coefficients.

Nitko and Feldt (1969) performed a computer simulation study of KR20 estimates and reported that, in contrast to general belief, the effect of item difficulty is minimal. Nitko (1968) employed the same method to investigate power functions for the test of significance of KR20 in one and two sample cases as proposed by

Feldt (1969). Weitzman (1967) reported the result of a simulation of test-retest reliability of a multi-choice test assuming a beta distribution of true scores. Shoemaker (1966) also used a computer simulation model to investigate the estimate of Cronbach's generalizability coefficient for unmatched data to clarify the extent to which stratification must be taken into account in the choice of the generalizability formula.

1.2.4 Summary

Recently, the concept of reliability has been modified and the restrictive classical assumptions of parallel tests relaxed substantially. However, the sampling distribution of reliability estimates based on Alpha or KR20 formulas are in general unknown except for the case when the unrealistic ANOVA model and underlying assumptions are used.

A number of fragmental attempts have been made recently to investigate the distribution by empirical methods, but there is no overall study into the statistical properties of the distribution under the more liberal concept of reliability either by analytical or empirical means.

The purpose of the present study is to investigate the statistical properties of the sampling distribution of reliability estimates when the classical parallel tests or more recent ANOVA models and the assumptions underlying these models are not all satisfied. More liberal concepts of reliability are to be examined in terms of models and assumptions underlying them, and sampling distributions under these models with various distributional assumptions

not necessary normal will be investigated by employing computer simulated statistical experiments. Comparisons are to be made of these results with those obtainable theoretically from the ANOVA model and normal theory to see the robustness of the theoretical distributions against the violation of assumptions.

1.3.0 Some Preliminary Specifications and Notations

1.3.1 Specifications

(a) With some exceptions, Greek letters will be used to denote population values, while the observations and sample quantities are denoted by Roman letters. To make notation simpler, no attempts are made to distinguish random variables from their observed values.

(b) Scalars will be denoted by capital and lower case letters, matrices will be denoted by underlined capital letters, and column vectors by underlined lower case letters. Row vectors will be indicated by transpose of column vectors, i.e., by priming them.

(c) An estimator of the population parameter and its value will be indicated by placing a caret or 'hat' over the parameter.

(d) The normal distribution with mean μ and variance σ^2 will be denoted by $N(\mu, \sigma^2)$. In general, a J-variate normal distribution having a mean vector $\underline{\mu}$ and a dispersion or variance-covariance matrix $\underline{\Sigma}$ will be denoted by $N(\underline{\mu}, \underline{\Sigma})$. The chi-square statistic with n degrees of freedom and the F-statistic with n and m degrees of freedom are denoted by χ^2_n and $F_{n;m}$ respectively.

(e) The expectation, and dispersion operations for a vector random variable $\underline{x}$ will be denoted by $E(\underline{x})$ and $D(\underline{x})$, namely,

$$E(\underline{x}) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \cdot \\ \cdot \\ \cdot \\ E(x_J) \end{bmatrix},$$

$$D(\underline{x}) = \begin{bmatrix} \text{Var}\ (x_1) & \text{Cov}\ (x_1,x_2) & \cdot & \cdot & \text{Cov}\ (x_1,x_J) \\ \text{Cov}\ (x_2,x_1) & \text{Var}\ (x_2) & \text{Cov}\ (x_2,x_3) & \cdot & \text{Cov}\ (x_2,x_J) \\ \text{Cov}\ (x_3,x_1) & \text{Cov}\ (x_3,x_2) & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \text{Cov}\ (x_J,x_1) & \cdot & \cdot & \cdot & \text{Var}\ (x_J) \end{bmatrix},$$

where $E(x_i)$, Var (x_i), and Cov (x_i,x_j) denote the usual expectation of x_i, variance of x_i, and covariance between x_i and x_j respectively.

(f) The correlation coefficient between two random variables x and y is denoted by Cor (x,y), namely,

$$\text{Cor}\ (x,y) = \frac{\text{Cov}\ (x,y)}{[\text{Var}\ (x)\ \text{Var}\ (y)]^{\frac{1}{2}}} .$$

(g) An identity matrix is denoted by $\underline{I}$, while a vector of length J whose elements are all 1's is denoted by $\underline{1}$.

(h) Dot subscripts are used to indicate sample means.

(g) Braces, { }, are used to indicate all elements in a set of variables.

1.3.2 Notations

The following is a brief glossary of important symbols used frequently.

i	indexing subscript for subjects in the sample, $i = 1,2,\ldots,I$
k	indexing subscript for subjects in population, $k = 1,2,\ldots$
j	indexing subscript for parts (items) of a test, $j = 1,2,\ldots,J$
I	sample size, a fixed constant
J	number of parts (items) in the test, a fixed constant
y_{ij}	the observed score random variable of subject i on the jth part test; it stands for the corresponding response strength variable for the binary item test
x_{ij}	the observed score random variable of subject i on the jth item, takes on values one or zero
τ_{ij}	the true score of subject i on the jth part
e_{ij}	the error score random variable of y_{ij}
m_i	the true score of subject i after adjustment is made for difference in difficulty levels among the J part tests
a_i	the effect or ability level of subject i in deviation form, $m_i-\mu$
β_j	the fixed effect of jth part, or the threshold constant for jth item; indicates the difficulty level of jth part (item)
μ	the expected value of m_i over the population
σ_A^2	the variance of m_i over the population, assumed to be common to all J parts under ETEM assumption
σ_{ej}^2	the variance of e_{ij} over the replications, assumed to be common to all subjects for all specific part j; defined in terms of response strength variables for the binary case

σ_e^2	common value of σ_{ej}^2 among the J parts under the homogeneity of error variance assumption
y_i	the unweighted sum of J part scores for subject i
x_i	the unweighted sum of J items for subject i
σ_j^2	the variance of the jth part (item) score
σ_y^2	the variance of y_i
σ_x^2	the variance of x_i
λ_j	the regression coefficient of y_{ij} on f_i under the uni-factor true score model, or the standard deviation of the true score of the jth part; the biserial correlation between x_{ij} and f_i for the binary item case
$\gamma_{jj'}$	the tetrachoric correlation between items j and j'
$\rho_{jj'}$	the inter item correlation coefficient between items j and j'
ε_{ij}	standardized error random variable, i.e., $e_{ij} = \sigma_{ej}\varepsilon_{ij}$
f_i	standardized true score random variable for continuous case i.e., $a_i = \sigma_A f_i$; for the binary item case the latent or factor score
$\underline{\Lambda}$	the factor loading matrix of size $J \times r$
r	the number of factors of the true score
$P_j(f)$	item characteristic function of the jth item
π_j	item difficulty of the jth item
$\phi(z)$	the normal density function
$\Phi(x)$	the cummulative normal distribution function
ρ_j	the reliability coefficient of the jth part
ρ	the reliability coefficient of the unweighted sum of J parts (items)

γ_A the kurtosis of true score a_i or f_i

γ_e the kurtosis of error score e_{ij} or ε_{ij}

γ_y the kurtosis of test score y_i

1.3.3 Vectors and Matrices

The following vectors and matrices are used frequently.

$$\underline{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ y_{iJ} \end{bmatrix}, \qquad \underline{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{iJ} \end{bmatrix}, \qquad \underline{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdot \\ \cdot \\ \cdot \\ \lambda_J \end{bmatrix} .$$

$$\underline{\Psi} = \begin{bmatrix} \sigma_{e1} & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \sigma_{e2} & 0 & \cdot & \cdot & 0 \\ 0 & 0 & \sigma_{e3} & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \sigma_{eJ} \end{bmatrix} ,$$

$$\underline{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdot & \cdot & \lambda_{1r} \\ \lambda_{21} & \lambda_{22} & \cdot & \cdot & \lambda_{2r} \\ \lambda_{31} & \lambda_{32} & \cdot & \cdot & \lambda_{3r} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \lambda_{J1} & \cdot & \cdot & \cdot & \lambda_{Jr} \end{bmatrix} ,$$

where λ_{jm} is the factor loading or regression coefficient of y_{ij} on the mth true or factor score under a multi-factor model.

1.3.4 Definitions

The following is a short list of definitions for the most often used terms in this paper.

ANOVA Model — Unless specified otherwise, this term refers to a two way mixed model analysis of variance linear model with one observation per cell. The levels of the row factor stand for the subjects in the sample, and the effects are assumed to be random, while the levels of the column factor stand for part tests or items of a composite test, and the effects are assumed to be fixed.

Parallel — Two measures are said to be classically or strictly parallel if (a) the test score may be considered consisting of two independent parts, true and error scores, (b) true scores are identical, and (c) error and total scores have idenical means and variances for each of the two measures.

Essentially Parallel (ANOVA) — The same as parallel measures except that the true scores may differ by a constant. The true, error and test scores have identical variances, but the means of the true scores may differ. Under the ANOVA model, the measurements are essentially parallel.

τ Equivalent	The same as parallel measures except that the variances of the error scores may differ. The variances of test scores may differ, but the means must be equal.
Essentially τ Equivalent (ETEM)	The same as τ equivalent measurements except that the true scores may differ by a constant. The variances of true scores must be identical, but means and variances of test scores may differ.
Congeneric True Score	The same as essentially τ equivalent measurements except that the true score is required only to measure a single trait. The true scores of two measures are linearly related, but their means and variances may differ.
Multi-factor (M.F.)	The same as congeneric true score case except that the tests measure more than one trait, i.e., the factorial structure of the true score could be more than one factor.

The above definitions of different but related types of measurements are compared in Table 1.1.

TABLE 1.1

Comparisons of the Definitions of Various Measures

Type of Measures	True Scores			Error Scores		Test Scores	
	Score[1]	Mean[2]	Var.[3]	Mean	Var.	Mean	Var.
Parallel	I[4]	I	I	0.0	I	I	I
Essentially Parallel (ANOVA)	D[5]	D	I	0.0	I	D	I
τ Equivalent	I	I	I	0.0	D	I	D
Essentially τ Equivalent (ETEM)	D	D	I	0.0	D	D	D
Congeneric	D	D	D	0.0	D	D	D
Multi-Factor (M.F.)	D	D	D	0.0	D	D	D

Note: [1]True scores for the same subject.

[2]Means in the population.

[3]Variances in the population.

[4]I: Identical among the measures.

[5]D: May differ among the measures.

CHAPTER TWO

TEST MODELS FOR THE CONTINUOUS PART SCORE CASES

Two distinct cases may be considered for a theory of reliability: the first is the case of continuous observed scores for the parts of a test, and the second is the case in which the scores of the parts are 'counter' or 'indicator' variables, i.e., a one is assigned for a correct response and zero for a wrong response. Due to the necessity of a different statistical treatment for each of the two cases, only the continuous case is duscussed in this chapter. The discussion is also limited to Type 1 sampling situations. The binary item situation will be discussed in the following chapter.

For the continuous score case, ANOVA type linear models are the most powerful and have a wide range of applicability. From among many possible models, the discussion is limited to a two way mixed model ANOVA with one observation per cell. Generalization to other more complex designs is a straight forward matter, however, complexity and difficulty of interpretation is a problem due to interaction effects.

2.1 ANOVA Model

A test consisting of J parts $(J \geq 2)$ is considered under the strict parallelism assumptions among the J part tests except that the means of the J parts may differ by a constant due to the difference in the difficulty levels of the parts. If the test is administered to a random sample of I subjects $(I \geq 2)$, the observed

score of the ith subject in the sample on the jth part, a random variable denoted by y_{ij}, may be written in a linear form in accordance with the classical theory of true and error scores, namely,

$$(2.1) \qquad y_{ij} = \tau_{ij} + e_{ij}, \quad i = 1,2,\ldots,I; \quad j = 1,2,\ldots,J,$$

where τ_{ij} and e_{ij} denote the true score and error score respectively.

An infinite idealized population of subjects denoted by $\mathcal{P}$, from which the sample of I subjects is supposedly drawn, is hypothesized and the findings on the sample are to be generalized to the population. Labelling the subjects in $\mathcal{P}$ as k $(k = 1,2,\ldots)$, the score y_{kj} may be conceptualized as the realization of a random process which may occur under repeated measurements on a single subject, labelled by k, on a fixed part test j with the assumption that the subject does not change or 'learn' over the repeated measurements, that is the replication is under experimentally independent conditions. Then the true score τ_{kj} may be considered the mean of y_{kj} over replications, or the expected value of y_{kj} over the distribution of y_{kj} for fixed k, or over the so-called 'propensity distribution' of y_{kj} (Lord and Novick, 1968, pp. 29-30). Mathematically τ_{kj} may be defined as the expectation of the random variable y_{kj}, for given k and j. The elements of y_{kj} have a joint distribution with respect to k in the population $\mathcal{P}$ and the number of replications.

By the assumption of parallelism, the true score τ_{kj} may be written,

$$(2.2) \qquad \tau_{kj} = m(k) + \beta_j ,$$

where β_j is a fixed constant specific to the jth part test representing the difficulty level of the part relative to the other parts, and $m(k)$ is the adjusted true score for subject k indicating the real ability level of the subject, and assumed to be independent of j. Thus, $m(k)$ may be considered a random variable with repsect to k distributed over the population $\mathcal{P}$.

Since the $\{\beta_j\}$ indicate only the relative difficulty levels among the J parts, without loss of generality it may be assumed that,

$$\sum \beta_j = 0 . \tag{2.3}$$

The labels of the subjects in the sample may be given by $\{k_1, k_2, \ldots, k_I\}$, and $m_i = m(k_i)$ where m_i is the adjusted true score of the ith subject in the sample. If μ and σ_A^2 denote the mean and variance of the adjusted true score $m(k)$, then they are the expected value and variance of the random variable $m(k)$ calculated with respect to the distribution of k in the population. Since each of the I subjects may also be considered as a randomly selected subject drawn from I identical populations with mean μ and variance σ_A^2, one subject chosen per population, each of the $\{m_i\}$ may also be considered as a random variable distributed independently and identically with expected value μ and variance σ_A^2.

The variance of the error random variable e_{kj}, calculated with respect to the propensity distribution of y_{kj}, for fixed k and j, shall be denoted by $\sigma_{ej}^2(k)$. Although it is conceivable that the brighter subjects with higher $m(k)$ might respond to the

test more consistently over replications, and have smaller variances, for the present discussion, it is assumed that $\sigma^2_{ej}(k)$ is the same for all subjects in the population and the common error variance is denoted by σ^2_{ej}, which depends only on j. This assumption is rather restrictive, but it is necessary since only one set of part test scores is assumed to be available for each subject in the sample, and therefore $\sigma^2_{ej}(k)$ would not be an observable quantity without this assumption.

Furthermore, following the assumptions of classical parallelism (e.g., Gulliksen, 1950, pp. 14-25), under the ANOVA model, it shall also be assumed that the error scores $\{e_{kj}\}$ have expected value zero and equal variance, denoted by σ^2_e, for all the J parts, i.e., homogeneity of error variance is also assumed among the part tests. In addition they are assumed to be independently and identically distributed, and independent of $\{m(k)\}$.

The effect of a subject labelled k in the population is defined as,

$$(2.4) \qquad a(k) = m(k) - \mu,$$

such that the effect of the ith subject in the sample, denoted by a_i, is

$$(2.5) \qquad a_i = m_i - \mu .$$

Applying (2.2) and (2.5), (2.1) becomes the basic model equation,

$$(2.6) \qquad y_{ij} = \mu + a_i + \beta_j + e_{ij} ,$$

with the following assumptions,

$$
(2.7) \quad \begin{cases} \text{(a)} & \{a_i\} \text{ and } \{e_{ij}\} \text{ are independent random variables,} \\ \text{(b)} & \sum \beta_j = 0, \\ \text{(c)} & \{a_i\} \text{ are identically distributed with } E(a_i) = 0, \\ & \text{and } \operatorname{Var}(a_i) = \sigma_A^2 \\ \text{(d)} & \{e_{ij}\} \text{ are identically distributed with } E(e_{ij}) = 0, \\ & \text{and } \operatorname{Var}(e_{ij}) = \sigma_e^2 \end{cases}
$$

Thus the expected value and variance of an observation y_{ij} is,

$$
(2.8) \qquad E(y_{ij}) = \mu + \beta_j ; \qquad \operatorname{Var}(y_{ij}) = \sigma_A^2 + \sigma_e^2 .
$$

If y_i denotes the unweighted sum of the J part scores for subject i, namely,

$$
(2.9) \qquad y_i = \sum_j y_{ij} = J\mu + Ja_i + \sum_j e_{ij} ,
$$

then,

$$
(2.10) \qquad E(y_i) = J\mu ; \qquad \operatorname{Var}(y_i) = J^2 \sigma_A^2 + J \sigma_e^2 .
$$

The reliability of a test is defined to be the ratio of the variance due to individual difference or the 'effect' of subjects to the total test score variance (Lord and Novick, 1968, p. 61). For a part j,

$$
(2.11) \qquad \rho_j = \frac{\operatorname{Var}(a_i)}{\operatorname{Var}(y_{ij})} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2} = \frac{\theta}{(1 + \theta)} ,
$$

where $\theta = \sigma_A^2/\sigma_e^2$ is the so-called signal-noise ratio or the square of sensitivity of a part test score (Jackson and Ferguson, 1941, p. 40).

For the total score,

$$(2.12) \qquad \rho = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2/J} = \frac{J\theta}{1 + J\theta} .$$

Because,

$$\begin{aligned} \text{Cov } (y_{ij}, y_{ij'}) &= E\{[y_{ij} - E(y_{ij})][(y_{ij'} - E(y_{ij'})]\} \\ &= E[(a_i + e_{ij})(a_i + e_{ij'})] \\ &= \sigma_A^2 , \end{aligned}$$

and the correlation coefficient between part j and j' $(j \neq j')$ is

$$\begin{aligned} \text{Cor } (y_{ij}, y_{ij'}) &= \text{Cov } (y_{ij}, y_{ij'})/[\text{Var } (y_{ij}) \text{ Var } (y_{ij'})]^{\frac{1}{2}} \\ &= \sigma_A^2/(\sigma_A^2 + \sigma_e^2) = \rho_j . \end{aligned}$$

The common reliability among the J parts, ρ_j, is the so-called 'intra-class' correlation coefficient which is the ordinary correlation coefficient between the part scores y_{ij} and $y_{ij'}$ under the ANOVA assumptions above. Or alternatively, ρ_j is the square of the index of reliability, which is the correlation between y_{ij} and a_i, since,

$$\begin{aligned} \text{Cov } (y_{ij}, a_i) &= E[(a_i + e_{ij})(a_i)] = \sigma_A^2 , \\ \text{Cor } (y_{ij}, a_i) &= \text{Cov } (y_{ij}, a_i)/[\text{Var } (y_{ij}), \text{Var } (a_i)]^{\frac{1}{2}} \\ &= \sigma_A/(\sigma_A^2 + \sigma_e^2)^{\frac{1}{2}} = \rho_j^{\frac{1}{2}} . \end{aligned}$$

Under this model no assumption of random sampling of parts is required. The models used by Hoyt (1941), Ebel (1951), Winer (1962, p. 124), Lord (1964), Feldt (1965), and Maguire and Hazlett (1969) are essentially the same as this, although some treat the fixed effect $\{\beta_j\}$ as random effects assuming the existence of a population of part tests and random sampling of J parts from it. As has been shown, the assumption is not necessary.

By the usual mathematical presentation (e.g., Scheffé, 1959, p. 261) the unbiased estimator of σ_A^2 and σ_e^2 are given by,

$$\hat{\sigma}_A^2 = (MS_A - MS_e)/J; \qquad \hat{\sigma}_e^2 = MS_e ,$$

where MS_A and MS_e are mean squares for subject effects and errors respectively. They are obtainable from an ANOVA table given as the following:

TABLE 2.1

Two Way Mixed Model ANOVA Table

Source	S.S.	D.F.	M.S.	E(M.S.)
Subject	$SS_A = J\sum(y_{i.}-y_{..})^2$	$I - 1$	$MS_A=SS_A/(I-1)$	$\sigma_e^2 + J\sigma_A^2$
Parts	$SS_B = I\sum(y_{.j}-y_{..})^2$	$J - 1$	$MS_B=SS_B/(J-1)$	$\sigma_e^2+I(\sum\beta_j^2)/(J-1)$
Errors	$SS_E = \sum_i\sum_j(y_{ij}-y_{i.}-y_{.j}+y_{..})^2$	$\nu = (I-1)(J-1)$	$MS_e = SS_e/\nu$	σ_e^2

Therefore if the reliability ρ_j or ρ is estimated by

substituting the unbiased estimator of variance components into (2.11) or (2.12),

$$(2.13)\quad \begin{cases} \text{(a)} & \hat{\rho}_j = \dfrac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2} = \dfrac{MS_A - MS_e}{MS_A + (J-1)MS_e} = \dfrac{F - 1}{F + J - 1} \\ \\ \text{(b)} & \hat{\rho} = \dfrac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2/J} = \dfrac{MS_A - MS_e}{MS_A} = 1 - 1/F \ , \end{cases}$$

where F is the ratio of mean squares, namely $F = MS_A/MS_e$.

The derivations up to and including equation (2.13) are valid without any distributional assumptions on $\{a_i\}$ and $\{e_{ij}\}$. In order to obtain a sampling distribution of the estimate (2.13), distributional assumptions are necessary. The simplest normal assumptions are

$$(2.14)\quad \begin{cases} \text{(a)} & \text{all } \{a_i\} \text{ are distributed as } N(0,\sigma_A^2) \ , \\ \\ \text{(b)} & \text{all } \{e_{ij}\} \text{ are distributed as } N(0,\sigma_e^2) \ . \end{cases}$$

With the above assumptions, model (2.6) is identical to the two way mixed model ANOVA with one observation per cell (Scheffé, 1959, p. 261). It can be shown that $SS_A/(J\sigma_A^2 + \sigma_e^2)$ and SS_e/σ_e^2 are distributed as chi-square with $I-1$ and ν degrees of freedom respectively, or

$$(2.15)\qquad SS_A = (J\sigma_A^2 + \sigma_e^2)\ \chi_{I-1}^2; \qquad SS_e = \sigma_e^2\ \chi_\nu^2 \ ,$$

hence, F is

$$(2.16) \qquad F = MS_A/MS_e = \frac{SS_A/(I-1)}{SS_e/\nu} = \frac{(J\sigma_A^2 + \sigma_e^2)\chi^2_{I-1}/(I-1)}{\sigma_e^2\ \chi_\nu/\nu} = (1+J\theta)F_{I-1;\nu} .$$

Therefore, from (2.13) and (2.16), the following relationship between F-statistic and ρ and $\hat{\rho}$, or ρ_j and $\hat{\rho}_j$ can be made:

$$(2.17) \qquad \begin{cases} \text{(a)} \qquad F_{I-1;\nu} = \dfrac{1-\rho}{1-\hat{\rho}} \\[2ex] \text{(b)} \qquad \qquad = \dfrac{[1+(J-1)\,\hat{\rho}_j][1-\rho_j]}{[1+(J-1)\,\rho_j][1-\hat{\rho}_j]} \end{cases} .$$

Feldt (1965), Nitko and Feldt (1969), Nitko (1968), and Cleary and Linn (1968) derived the above formula, and even applied it to the sampling distribution of KR20 estimates. Kristof (1963) obtained the same results by means of maximum likelihood methods using a multi-normal assumption. He obtained an estimate of intra-class correlation coefficient, which is equal to $\hat{\rho}_j$, and gave the estimate of the reliability of the total $\hat{\rho}$, called a step-up reliability, by using the general Spearman-Brown formula. However, this result is not new for mathematical statisticians. For example Scheffé gave similar results (1959, pp. 226-229).

Because (2.17) gives the relationship between the sample statistic $\hat{\rho}_j$, or $\hat{\rho}$ and the population parameter ρ_j or ρ in terms of the well-known F-statistic, the sampling distribution of reliability estimates can be determined; thus, it is possible to make inferences about the reliability, and to calculate confidence intervals. Within the essentially parallel assumptions, the sampling distribution of the reliability would not raise any questions provided the

assumptions (2.7) and (2.14) are all met and the model as given by equation (2.6) is adequate.

As a special case of the model, let all of the fixed effects $\{\beta_j\}$ be equal to zero, then the model (2.6) reduces to

$$(2.18) \qquad y_{ij} = \mu + a_i + e_{ij}, \quad i = 1,2,\ldots,I; \quad j = 1,2,\ldots,J .$$

This model is identical to the one way random effect model ANOVA (Scheffé, 1959, pp. 221-235), and it can be shown that all the formulas given above are valid with SS_e and ν replaced by $(SS_B + SS_e)$ and $I(J - 1)$, and MS_e modified accordingly. Model (2.18) is equivalent to the classical parallelism assumptions (e.g., Gulliksen, 1950, p. 11) except for the distributional assumptions which are not required under the classical test theory. Kristof's case 2 and Maguire and Hazlett's case C (1969) correspond to this model

Because the variance of random variables $\{a_i\}$ and $\{e_{ij}\}$ are equal to σ_A^2 and σ_e^2 respectively under the ANOVA model, they may be rewritten in terms of standard random variables $\{f_i\}$ and $\{\varepsilon_{ij}\}$, namely,

$$a_i = \sigma_A f_i, \quad i = 1,2,\ldots,I, \quad \text{and}$$

$$e_{ij} = \sigma_e \varepsilon_{ij}, \quad i = 1,2,\ldots,I; \quad j = 1,2,\ldots,J .$$

Then the model equation (2.6) becomes

$$y_{ij} = \mu + \sigma_A f_i + \beta_j + \sigma_e \varepsilon_{ij}, \quad i = 1,2,\ldots,I; \quad j = 1,2,\ldots,J .$$

The above equations can be rewritten in a matrix equation

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ y_{iJ} \end{bmatrix} = \begin{bmatrix} \mu+\beta_1 \\ \mu+\beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu+\beta_J \end{bmatrix} + \begin{bmatrix} \sigma_A \\ \sigma_A \\ \cdot \\ \cdot \\ \cdot \\ \sigma_A \end{bmatrix} \cdot \begin{bmatrix} f_i \end{bmatrix} + \begin{bmatrix} \sigma_e & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \sigma_e & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & \sigma_e \end{bmatrix} \cdot \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \\ \\ \\ \varepsilon_{iJ} \end{bmatrix},$$

or, using the notations given in Section 1.3 of Chapter One,

(2.6') $\qquad \underline{y}_i = \underline{\mu} + \underline{\lambda}\, f_i + \underline{\Psi}\, \underline{\varepsilon}_i , \qquad i = 1,2,\ldots,I$,

with the limitations $\lambda_1 = \lambda_2 = \ldots = \lambda_J = \sigma_A$, and $\Psi_{11} = \Psi_{22} = \ldots = \Psi_{JJ} = \sigma_e$. The assumptions of (2.7) may be rewritten as,

(2.7')

(a) all $\{f_i\}$ and $\{\varepsilon_{ij}\}$ are independent random variables,

(b) $\sum \mu_j = \mu$, where $\mu_j = \mu + \beta_j$,

(c) all $\{f_i\}$ are identically distributed with $E(f_i) = 0$, $\mathrm{Var}\,(f_i) = 1$, or $E(\underline{\lambda}\, f_i) = \underline{0}$, $D(\underline{\lambda}\, f_i) = \underline{\lambda}\,\underline{\lambda}'$,

(d) all $\{\underline{\varepsilon}_i\}$ are identically distributed with $E(\underline{\varepsilon}_i) = \underline{0}$, $D(\underline{\varepsilon}_i) = \underline{I}$, or $E(\underline{\Psi}\,\underline{\varepsilon}_i) = \underline{0}$, $D(\underline{\Psi}\,\underline{\varepsilon}_i) = \underline{\Psi}^2$.

The distributional assumption of (2.14) becomes

(2.14')

(a) all $\{f_i\}$ are $N(0,1)$, or all $\{\underline{\lambda}\, f_i\}$ are $N(\underline{0}, \underline{\lambda}\,\underline{\lambda}')$,

(b) all $\{\underline{\varepsilon}_i\}$ are $N(\underline{0}, \underline{I})$, or all $\{\underline{\Psi}\,\underline{\varepsilon}_i\}$ are $N(\underline{0}, \underline{\Psi}^2)$.

2.2 Essentially τ Equivalent Measurements Model

Under the ANOVA model, $\tau_{ij} = \mu + a_i + \beta_j$, hence with $j \neq j'$, $\tau_{ij} - \tau_{ij'} = \beta_j - \beta_{j'} = c$ where c is a constant which depends only on j and j'. Therefore, the part tests satisfy the conditons of the so-called essentially τ equivalent measurements (Lord and Novick, 1968, p. 50) which will be denoted as ETEM henceforth. Because the assumption of homogeneity of error variances is not required for the definition of ETEM, the error variance σ_{ej}^2 may depend on a specific j. Thus, assumption (d) of (2.7) may be modified to become,

(2.19) (d) $\{e_{ij}\}$ are distributed with $E(e_{ij}) = 0$; $\text{Var}\ (e_{ij}) = \sigma_{ej}^2$.

The variance of y_{ij} and the covariance between y_{ij} and $y_{ij'}$ are given by

$$(2.20)\qquad \begin{cases} \text{Var}\ (y_{ij}) = E[y_{ij} - E(y_{ij})]^2 = \sigma_A^2 + \sigma_{ej}^2\ , \\ \\ \text{Cov}\ (y_{ij}, y_{ij'}) = E\{[y_{ij} - E(y_{ij})][y_{ij'} - E(y_{ij'})]\} = \sigma_A^2\ . \end{cases}$$

Therefore, the reliability of jth part test is given by

$$(2.21)\qquad \rho_j = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_{ej}^2}\ ,$$

i.e., it depends on j, and hence, in general, under the ETEM model, there is not a common correlation coefficient among the J parts. Therefore the reliability of a part cannot be interpreted as an

intra-class correlation coefficient. The correlation coefficient between j and j' depends on j and j', because,

$$\mathrm{Cor}\ (y_{ij}, y_{ij'}) = \frac{\mathrm{Cov}\ (y_{ij}, y_{ij'})}{[\mathrm{Var}\ (y_{ij}) \mathrm{Var}(y_{ij'})]^{\frac{1}{2}}} = \frac{\sigma_A^2}{[(\sigma_A^2 + \sigma_{ej}^2)(\sigma_A^2 + \sigma_{ej'}^2)]^{\frac{1}{2}}} .$$

The reliability of the total test is given by,

$$\rho = \frac{\mathrm{Var}\ (J\ a_i)}{\mathrm{Var}\ (y_i)} = \frac{J^2\ \sigma_A^2}{J^2\ \sigma_A^2 + \sum \sigma_{ej}^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_{e.}^2/J} , \tag{2.22}$$

where $\sigma_{e.}^2$ is average of σ_{ej}^2 , i.e., $\sigma_{e.}^2 = (\sum \sigma_{ej}^2)/J$.

If σ_j^2 denotes the total variance of jth part test given by (2.20), the total test variance denoted by σ_y^2 is

$$\sigma_y^2 = \mathrm{Var}\ (\sum_j y_{ij}) = \sum_j \mathrm{Var}\ (y_{ij}) + \sum_{j \neq j'}\sum \mathrm{Cov}\ (y_{ij}, y_{ij'})$$

$$= \sum \sigma_j^2 + J(J-1)\ \sigma_A^2 .$$

Substituting this into (2.22)

$$\rho = \frac{J^2\ \sigma_A^2}{\sigma_y^2} = \frac{J}{J-1} [1 - \frac{\sum \sigma_j^2}{\sigma_y^2}] .$$

which is the well-known formula for the Alpha coefficient.

Novick and Lewis (1967) have shown that Alpha is equal to reliability ρ if and only if the ETEM assumption is satisfied. Otherwise, Alpha is, in general, lower than the reliability, namely

$$\text{(2.23)} \qquad \text{Alpha} \leq \rho \; .$$

The equality holds only if the ETEM assumption is true.

Alpha is usually estimated by

$$\text{(2.24)} \qquad \text{Alpha} = \frac{J}{J-1}\left[1 - \frac{\sum s_j^2}{s_y^2}\right] ,$$

where s_j^2 and s_y^2 are the usual sample variance of part test and the total test respectively. Kristof (1970) investigated the sampling distribution of the Alpha estimate for the case of $J = 2$, and showed that the distribution can be reduced to a Student's t-statistic by the maximum likelihood method. The sampling distribution for the general case is not yet known.

Classically, Alpha as a reliability is derived (e.g., Gulliksen, 1950, p. 223) by considering two J-parts tests that are parallel part by part, and then introducing the assumption that the covariance of a part in one test with the parallel part of the second test is equal in the average to the covariance between any two of the J parts within a test. If y_{ij}^* denotes the score of the jth part of the second test the assumption is,

$$\text{(2.25)} \qquad \sum_j \text{Cov}\,(y_{ij}, y_{ij}^*) = [\sum_{j \neq j'}\sum \text{Cov}\,(y_{ij}, y_{ij'})]/(J-1) \; .$$

Lord and Novick (1968, p. 92) have shown that the above assumption is satisfied if and only if the j parts are ETEM.

Under the ETEM assumption, the matrix model equation is the same as (2.6') except that the diagonal elements of $\underline{\Psi}$ may differ,

namely,

$$\underline{y}_i = \underline{\mu} + \underline{\lambda}\, f_i + \underline{\Psi}\, \underline{\varepsilon}_i \, ,$$

$$\Psi_{11} = \sigma_{e1}, \quad \Psi_{22} = \sigma_{e2}, \ldots, \Psi_{JJ} = \sigma_{eJ} \quad .$$

All the assumptions of (2.7') and (2.14') may be applied.

2.3 Congeneric True Score Model

Under the ETEM model, including the ANOVA model as a special case, the true score variance, σ_A^2, is assumed to be common to all J part tests. However, for some tests, it might be more reasonable to expect that the true score variance would depend on j, i.e., some of the part tests might discriminate better and have greater true score variances. Under this situation, the classical parallelism or ETEM assumption is no longer valid. Nevertheless, the model given by equation (2.6') may still be used by removing the restriction of equal $\{\lambda_j\}$, namely the elements of the vector $\underline{\lambda}$ may differ. The constant λ_j may be interpreted as a regression coefficient of y_{ij} on the standard true score f_i, or standard deviation of the jth true score. In scalar form the equation is

$$y_{ij} = \mu + \lambda_j f_i + \beta_j + \sigma_{ej}\, \varepsilon_{ij} \quad . \tag{2.26}$$

Under this model, the reliability is

$$(2.27)\quad \begin{cases} \text{(a)} \quad \rho_j = \dfrac{\mathrm{Var}\ (\lambda_j f_i)}{\mathrm{Var}\ (y_{ij})} = \dfrac{\lambda_j^2}{\lambda_j^2 + \sigma_{ej}^2} \\[2ex] \text{(b)} \quad \rho = \dfrac{\mathrm{Var}\ (\underline{1}'\ \underline{\lambda}\ f_i)}{\mathrm{Var}\ (\underline{1}'\ \underline{y}_i)} = \dfrac{\underline{1}'\ \underline{\lambda}\ \underline{\lambda}'\ \underline{1}}{\underline{1}'(\underline{\lambda}\ \underline{\lambda}' + \underline{\psi}^2)\underline{1}} \end{cases},$$

where $\underline{1}$ is a $J \times 1$ vector whose elements are all 1's. Since the ETEM assumption is not satisfied, in general, Alpha $\leq \rho$.

No formula is yet available for the direct estimation of the reliability under this model, hence the estimate of Alpha is generally used as an estimate of the lower bound for the reliability. Cronbach, Ikeda and Avner (1964) used a similar model in their effort to approximate the generalizability coefficient by an intra-class correlation coefficient. However their model, which involves sampling of part tests (Type 2 sampling), assumes a uniform distribution of λ^2, unlike the present model where the $\{\lambda_j\}$ are assumed to be fixed constants. Jöreskog (1968, 1970) named this model as the congeneric test model.

2.4 Multi-Factor True Score Model

The three models reviewed in the previous sections implicitly assumed that the test measures only one ability or trait, represented by f_i, i.e., it is assumed that the factorial structure of the true score is a uni-factor model. However, for certain types of tests, the assumption is too restrictive, and a more general model which would accommodate more than one true score structure is desirable.

If $\underline{\lambda}$ and f_i are replaced by a $J \times r$, $(1 \leq r < J)$ constant factor loading matrix $\underline{\Lambda}$ and a $r \times 1$ standard random factor score vector $\underline{f}_i$ respectively, the model of equation (2.6'), becomes the well-known multi-factor model, (e.g., Browne, 1969; Jöreskog, 1970), namely,

(2.28) $$\underline{y}_i = \underline{\mu} + \underline{\Lambda}\,\underline{f}_i + \underline{\Psi}\,\underline{\varepsilon}_i \ ,$$

with,

$$E(\underline{y}_i) = \underline{\mu} \ ; \qquad D(\underline{y}_i) = \underline{\Lambda}\,\underline{\Lambda}' + \underline{\Psi}^2 \ .$$

Therefore, the reliability of the total test, ρ, is given by,

(2.29) $$\rho = \frac{\mathrm{Var}\,(\underline{1}'\,\underline{\Lambda}\,\underline{f}_i)}{\mathrm{Var}\,(\underline{1}'\,\underline{y}_i)} = \frac{\underline{1}'\,\underline{\Lambda}\,\underline{\Lambda}'\,\underline{1}}{\underline{1}'(\underline{\Lambda}\,\underline{\Lambda}' + \underline{\Psi}^2)\underline{1}} \ .$$

If the estimate $\hat{\underline{\Lambda}}$ is available, an estimate of the reliability would be,

(2.30) $$\hat{\rho} = \frac{\underline{1}'\,\hat{\underline{\Lambda}}\,\hat{\underline{\Lambda}}'\,\underline{1}}{s_y^2} \ .$$

The statistical properties of this statistic are unknown, and there is no agreed upon mean to obtain estimates of the factor loading matrix.

Under this model Alpha is in general the lower bound for the reliability as with the congeneric model. The equality is true if and only if the parts are ETEM, i.e., $r = 1$. For this case the factor loading matrix $\underline{\Lambda}$ becomes the vector $\underline{\lambda}$ with all elements equal, and

the standard deviations of the true scores are equal, as $\lambda_j = \sigma_A$ for all $j = 1,2,\ldots,J$. If the error variances are all equal among the J part tests, the model becomes identical to the ANOVA model. Therefore the multi-factor model equation (2.28) includes the ANOVA, ETEM and the congeneric model as special cases.

Under this general model the assumptions are

(2.31)

(a) all $\{\underline{f}_i\}$ and $\{\underline{\varepsilon}_i\}$ are independent random vector variables,

(b) $\sum \mu_j = \mu$, where $\mu_j = \mu + \beta_j$,

(c) all $\{\underline{f}_i\}$ are identically distributed with $E(\underline{f}_i) = \underline{0}$, $D(\underline{f}_i) = \underline{I}$, or $E(\underline{\Lambda}\ \underline{f}_i) = \underline{0}$ and $D(\underline{\Lambda}\ \underline{f}_i) = \underline{\Lambda}\ \underline{\Lambda}'$,

(d) all $\{\underline{\varepsilon}_i\}$ are identically distributed with $E(\underline{\varepsilon}_i) = \underline{0}$, and $D(\underline{\varepsilon}_i) = \underline{I}$, or $E(\underline{\Psi}\ \underline{\varepsilon}_i) = \underline{0}$, and $D(\underline{\Psi}\ \underline{\varepsilon}_i) = \underline{\Psi}^2$.

The normality assumption becomes

(2.32)

(a) all $\{\underline{f}_i\}$ are distributed as $N(\underline{0}, \underline{I})$, or $\{\underline{\Lambda}\ \underline{f}_i\}$ are $N(\underline{0}, \underline{\Lambda}\ \underline{\Lambda}')$,

(b) all $\{\underline{\varepsilon}_i\}$ are distributed as $N(\underline{0}, \underline{I})$, or $\{\underline{\Psi}\ \underline{\varepsilon}_i\}$ are $N(\underline{0}, \underline{\Psi}^2)$.

2.5 Summary

Four basic models which might be used for simulation of test scores are examined in this chapter under the assumption that a test has been split into J parts whose scores are continuous random variables.

The most general model is found to be the multi-factor model. The other three models are special cases of this model with additional assumptions or restrictions on the parameters.

With uni-factor assumptions, i.e., $r = 1$, the congeneric model is the most general one, which includes the other two models as special cases. However, the Alpha coefficient is identical to the reliability of the total test score if and only if the ETEM assumption is satisfied. Hence under the multi-factor or congeneric model, in general, the Alpha coefficient is a lower bound for the reliability.

With the homogeneity of error variance assumption the ETEM model becomes identical to the ANOVA model, the most restrictive one, and the distribution of reliability estimate is related to an F-statistic. Under more general models, the distribution is in general unknown.

If equal means are assumed among the J parts, the ANOVA model becomes identical with the classical parallelism model except for the distributional assumptions.

CHAPTER THREE

TEST MODEL FOR THE BINARY ITEM SCORE CASE

For a test consisting of J binary items as the parts of the test, the Kuder-Richardson formula 20(KR20) has been widely used as a special case of the Alpha coefficient with little investigation of its statistical properties. Feldt (1965), and Cleary and Linn (1968) treated the discrete case as a continuous part score case. However, the imposition of the zero-one scoring scheme violates not only the assumption of continuity of part scores, but also homogeneity of error variances and independence of true and error scores. The violation of the assumptions of the ANOVA model was fully discussed by Feldt (1965).

3.1 Normal Ogive Model

To investigate the statistical properties of test scores of binary item tests, a number of mathematical models have been proposed such as the normal ogive, logistic, and binomial models. The first two assume existence of a latent trait or factor score f, which can account for the subjects behavior or performance. The binomial model relies on the 'strong true score' theory (Lord, 1965; Birnbaum, 1968, pp. 508-529). In this model the conditional distribution of the test score for a given true score is assumed to be binomial.

In the following, the discussion is restricted to the statistical properties of reliability and KR20 under the normal ogive model. Extensions to other models may be done in a similar way. Although

the multi-factor model for the binary test is also possible, for the sake of simplicity, only the uni-factor case will be examined. Under this model, the random variable representing the latent trait or factor scores, f, is assumed to be independently distributed as $N(0,1)$, for all subjects in the population P, as under the continuous part test score models. It is also assumed that the response of the ith subject, with latent trait $f = f_i$ to each of J items, is determined by a hypothetical intervening random variable y_{ij} which shall be called the 'response strength variable' according to Bock and Liberman (1970). Since only the relative strength of y_{ij} is of interest, without loss of generality, it may be assumed that y_{ij} is distributed with expected value zero and unit variance, i.e., it is a standard random variable. In addition y_{ij} is assumed to be subject to random error, and if the value of y_{ij} for the ith subject on the jth item exceeds a certain threshold constant specific to the item, denoted by β_j, the observed score of the subject, denoted by x_{ij} is equal to one, otherwise it is equal to zero. In this case the continuous response strength variable y_{ij} may be written as a linear congeneric true score model noted in Section 2.3 of Chapter Two.

$$(3.1) \qquad y_{ij} = \lambda_j f_i + \sigma_{ej}\, \varepsilon_{ij}, \quad i = 1,2,\ldots,I; \quad j = 1,2,\ldots,J,$$

where λ_j is a constant regression coefficient specific to item j, σ_{ej} is the standard deviation of the error scores for jth item, while ε_{ij} is a standard random variable for errors as before.

In vector notation,

$$(3.2) \qquad \underline{y}_i = \underline{\lambda}\, f_i + \underline{\Psi}\, \underline{\varepsilon}_i \;,$$

where $\underline{\lambda}$, $\underline{\Psi}$, and $\underline{\varepsilon}_i$ are as defined for the congeneric model, except that the continuous part tests are replaced by dichotomous items. Also $D(\underline{y}_i) = \underline{\lambda}\,\underline{\lambda}' + \underline{\Psi}^2$ as before, and the diagonal elements of $D(\underline{y}_i)$ are the variances of y_{ij}, and are assumed to be unity, i.e., $1 = \lambda_j^2 + \sigma_{ej}^2$ for all $j = 1,\ldots,J$.

Thus model equation (3.1) may be rewritten,

$$(3.3) \qquad y_{ij} = \lambda_j f_i + (1-\lambda_j^2)^{\frac{1}{2}} \varepsilon_{ij}, \qquad i = 1,2,\ldots,I; \qquad j = 1,2,\ldots,J,$$

where the standard random variables $\{\varepsilon_{ij}\}$ are assumed to be distributed independently as $N(0,1)$.

The constant λ_j may also be interpreted as the index of reliability of the jth response strength variable, since

$$\begin{aligned} \mathrm{Cor}\ (y_{ij}, f_i) &= \mathrm{Cov}\ (y_{ij}, f_i) \\ &= E\{[\lambda_j f_i + (1-\lambda_j^2)^{\frac{1}{2}} \varepsilon_{ij}] f_i\} \\ &= \lambda_j \ . \end{aligned}$$

By definition the correlation coefficient between y_{ij} and f_i is equal to the biserial correlation coefficient between x_{ij} and f_i, therefore, λ_j is actually the biserial correlation between the latent trait variable f and observable item score x_{ij}. Since the correlation between y_{ij} and the individual effect or the true score $\lambda_j f_i$ is λ_j, the square of λ_j may also be interpreted as the reliability of the jth response strength variable.

3.2 Item Parameters

The jth item characteristic function $P_j(f)$ is defined to be the expected value of x_{ij} given f for subject i, (e.g., Lord and Novick, 1968, p. 360), namely,

$$P_j(f) = E(x_{ij} \mid f = f_i) = \text{Probability } (x_{ij} = 1 \mid f = f_i) \ . \tag{3.4}$$

Lord (1952, 1953), Lord and Novick (1968, pp. 358-394), Samejima (1969), Bock and Liberman (1970) and many others have investigated the item characteristic function under the normal ogive model.

The expected value and variance of response strength variable y_{ij} given a fixed subject with $f = f_i$, are given as,

$$E(y_{ij} \mid f = f_i) = \lambda_j f_i ; \quad \text{Var } (y_{ij} \mid f = f_i) = 1 - \lambda_j^2 \ . \tag{3.5}$$

The distribution of y_{ij} for fixed $f = f_i$ is normal with expected value $\lambda_j f_i$ and variance $1 - \lambda_j^2$, or $N[\lambda_j f_i, (1 - \lambda_j^2)]$. Thus the probability that subject i with the latent trait $f = f_i$ will respond correctly to item j, as indicated by observed value $x_{ij} = 1$, is

$$P_j(f) = \text{Probability } (x_{ij} = 1 \mid f = f_i)$$

$$= \frac{1}{[2\pi(1 - \lambda_j^2)]^{\frac{1}{2}}} \int_{\beta_j}^{\infty} \text{Exp } \frac{-(y_{ij} - \lambda_j f)^2}{2(1 - \lambda_j^2)} \, dy_{ij} \ .$$

Applying the transformation $z = (y_{ij} - \lambda_j f)/(1 - \lambda_j^2)^{\frac{1}{2}}$, $P_j(f)$ becomes

$$P_j(f) = \int_{g_j}^{\infty} \phi(z)dz = \Phi(-g_j) \ , \tag{3.6}$$

where $\phi(z)$ and $\Phi(-g)$ are the respective standard normal density and distribution functions. The value $g_j(f)$ is given by,

$$(3.7) \qquad g_j(f) = -(\lambda_j f - \beta_j)/(1 - \lambda_j^2)^{\frac{1}{2}} .$$

Using generally accepted notation (e.g., Lord and Novick, 1968), $g_j(f)$ may be rewritten as,

$$(3.8) \qquad g_j(f) = -a_j(f - b_j) \quad ,$$

hence

$$a_j = \lambda_j/(1 - \lambda_j^2)^{\frac{1}{2}} \quad ,$$

and

$$(3.9) \qquad \begin{cases} b_j = \beta_j/\lambda_j \\ \beta_j = a_j \; b_j/(1 + a_j^2)^{\frac{1}{2}} . \end{cases}$$

The item parameters a_j and b_j have been referred to as item 'discrimination power' and 'difficulty index' by Lord and Novick (1968. pp. 368-368).

The difficulty of item j is defined as expected value of x_{ij}, namely,

$$(3.10) \qquad \pi_j = \text{Probability } (x_{ij} = 1) = E(x_{ij}) = \int_{-\infty}^{\infty} P_j(f) \; \phi(f)df .$$

After some algebraic manipulation, it can be shown (Lord and Novick, 1968, p. 337) that, $\pi_j = \Phi(-\beta_j)$.

Since $E(x_{ij}^2) = E(x_{ij})$, the variance of jth item is given by,

$$(3.11) \quad \sigma_j^2 = \mathrm{Var}\ (x_{ij}) = E(x_{ij}^2) - [E(x_{ij})]^2 = \pi_j - \pi_j^2 = \pi_j(1 - \pi_j) \ .$$

3.3 Reliability of Binary Item Test

Since direct decomposition of the response score x_{ij} into independent true and error scores is impossible for the binary item scores, there is no direct way of obtaining the variance ratio of true score variance to total test score variance which has been defined as the reliability of a test. Nevertheless the population reliability may be obtained by resorting to the correlation method, namely, by calculating the correlation coefficient between

$$x_i = \sum_j x_{ij} \qquad \text{and} \qquad x_i^* = \sum_j x_{ij}^* \ ,$$

where x_{ij}^* is the score of a hypothetical test item which is parallel in the classical sense to the jth item of the test.

Then,

$$(3.12) \quad \rho = \mathrm{Cor}\ (\sum_j x_{ij}, \sum_j x_{ij}^*) = \frac{\mathrm{Cov}\ (\sum_j x_{ij}, \sum_j x_{ij}^*)}{[\mathrm{Var}\ (x_i)\ \mathrm{Var}\ (x_i^*)]^{\frac{1}{2}}}$$

$$= \frac{\sum_j \sum_{j*} \sigma_j \sigma_{j*} \rho_{jj*}}{\sigma_x^2} \ ,$$

where ρ_{jj*} is the inter-item correlation between item j and j*, σ_j^2 is the variance of the jth item, and σ_x^2 is the variance of the

total test score x_i. The test variance σ_x^2 may be given in terms of inter-item correlation and item variance, namely,

$$(3.13) \qquad \mathrm{Var}\,(x_i) = \mathrm{Var}\,(\sum_j x_{ij}) = \sum_j \mathrm{Var}\,(x_{ij}) + \sum_{j \neq j'}\sum \mathrm{Cov}\,(x_{ij}, x_{ij'})$$

$$= \sum \sigma_j^2 + \sum_{j \neq j'}\sum \sigma_j\, \sigma_{j'}\, \rho_{jj'} \quad .$$

To obtain ρ and σ_x^2, the inter-item covariance $\sigma_j\, \sigma_{j'}\, \rho_{jj'}$ must be evaluated in terms of the item parameters λ_j and $\lambda_{j'}$. Lord and Novick (1968, p. 379) showed that for any two items j and j', the tetrachoric correlation between x_{ij} and $x_{ij'}$, denoted by $\gamma_{jj'}$, can be expressed as the product of the two biserial correlations λ_j and $\lambda_{j'}$ by performing integration of the tri-variable distribution x_{ij}, $x_{ij'}$, and f, namely

$$(3.14) \qquad \gamma_{jj'} = \lambda_j\, \lambda_{j'} \quad .$$

It can be shown (e.g., Kendall and Stuart, 1963, p. 161 and 1967, p. 306) that the inter-item covariance may be expressed as an infinite power series of $\gamma_{jj'}$ using Tchebycheff-Hermite polynomials, denoted by $H_n(\beta)$ (Kendall and Stuart, 1963, p. 155). Then,

$$(3.15) \qquad \mathrm{Cov}\,(x_{ij}, x_{ij'}) = \sigma_j\, \sigma_{j'}\, \rho_{jj'}$$

$$= [\phi(\beta_j)\phi(\beta_{j'})][\gamma_{jj'} + 0.5\, \beta_j \beta_{j'}\, \gamma_{jj'}^2 \ldots]$$

$$= [\phi(\beta_j)\phi(\beta_{j'})] \sum_{n=1}^{\infty} [H_{n-1}(\beta_j)\, H_{n-1}(\beta_{j'}) \gamma_{jj'}^n]/n! \quad .$$

Therefore, the covariance may be calculated numerically.

By the results of equations (3.11), (3.13), (3.14), and (3.15), the reliability ρ, given by equation (3.12) may be evaluated numerically if the item parameters $\{\beta_j\}$ and $\{\lambda_j\}$ are specified.

3.4 KR20 Coefficient and Its Estimate

The Alpha coefficient for the binary item test, KR20, is defined as,

$$(3.16) \qquad KR20 = \frac{J}{J-1}\left[1 - \frac{\sum \sigma_j^2}{\sigma_x^2}\right] = \frac{J}{J-1}\left[1 - \frac{\sum \pi_j (1 - \pi_j)}{\sigma_x^2}\right]$$

which is equal to the reliability ρ if and only if the ETEM condition of (2.25) is satisfied. In general the condition is not satisfied, hence,

$$(3.17) \qquad KR20 \leq \rho \quad .$$

Using the results of equations (3.10), (3.13), (3.15), and (3.16), KR20 may also be evaluated numerically if the item difficulty $\{\pi_j\}$ and biserial correlation $\{\lambda_j\}$ are specified, provided f_i and $\{\varepsilon_{ij}\}$ are distributed independently as $N(0,1)$.

The ETEM condition of (2.25) for binary item cases in terms of the powe series of (3.15) may be written as,

$$\sum_j [\phi(\beta_j)^2 [\sum_{n=1}^{\infty} (H_{n-1}^2(\beta_j)\lambda_j^{2n})/n!]$$

$$= \{\sum_{j \neq j'}\sum \phi(\beta_j)\phi(\beta_{j'}) \sum_{n=1}^{\infty} [H_{n-1}(\beta_j)\, H_{n-1}(\beta_{j'})\gamma_{jj'}^n]/n!\}/(J-1) \quad ,$$

which is true if $\beta_1 = \beta_2 = \ldots = \beta_J$ and $\lambda_1 = \lambda_2 = \ldots = \lambda_J$. This means that all the item parameters are equal.

An estimate of KR20 is obtained by substitutiing the sample estimates of $\{\pi_j\}$ and σ_x^2 , namely,

$$(3.18) \qquad \widehat{KR20} = \frac{J}{J-1} \times [1 - \frac{\sum \hat{\pi}_j(1 - \hat{\pi}_j)}{s_x^2}],$$

where $\hat{\pi}_j$ is the sample difficulty of the jth item, and s_x^2 is the sample variance of the test score x_i, given by,

$$(3.19) \qquad \begin{cases} \hat{\pi}_j = (\sum_i x_{ij})/I , \\ s_x^2 = [\sum_i (x_i - x_.)^2]/I . \end{cases}$$

Unlike the case of reliability estimates under the ANOVA model, the exact sampling distribution of KR20 estimates given by (3.18) is unknown even with the restrictive mathematical models and assumptions. Aoyama's formulas (1957), provided an approximation for the expected value and the variance of KR20 estimates without any distributional assumptions. He gave approximate formulas for $E(\widehat{KR20})$ and $Var\ (\widehat{KR20})$ as,

$$(3.20) \qquad E(\widehat{KR20}) = KR20 + 0(1/J) ,$$

and

$$(3.21) \qquad Var\ (\widehat{KR20}) \leq \frac{1}{I(J-1)^2} (\frac{I}{I-1})^2 (\delta_2 + 15 + \frac{14}{J} + \frac{18}{J^2} x_m^2) ,$$

where $0(1/J)$ is a term of the order of $1/J$, δ_2 is the kurtosis of

the distribution of x_i and x_m is the minimum score. Formula (3.20) indicates the estimate is biased, while (3.21) suggests a bound for the standard error of KR20 estimates, but is of little use since it involves the unknown parameter δ_2, which would be very difficult if not impossible to evaluate.

3.5 Summary

To examine the feasibility of simulating binary item test scores, the well-known normal ogive model is reviewed in terms of two basic item parameters, namely the item difficulty $\{\pi_j\}$, and the biserial correlation between items and the latent trait or factor score f_i, i.e., $\{\lambda_j\}$.

With the help of the 'response strength variable', a model equation similar to (2.26) used in the previous chapter was introduced.

Item characteristic functions and item indices are also examined in terms of the two sets of parameters.

The population reliability and KR20 were found to be amenable to calculation through numerical means in terms of these parameters. The exact sampling distribution of KR20 estimates is in general unknown.

CHAPTER FOUR

RATIONALE FOR SIMULATION, COMPUTER PROGRAMS, AND METHOD OF INVESTIGATION

4.1 Violation of ANOVA Model and Assumptions

The review in the previous three chapters indicated that the exact sampling distribution of the reliability estimates is largely unknown except for the special case of the ANOVA model under rather restrictive assumptions. The distributional thoery and inferences based on this model are valid, if and only if all of the underlying assumptions of (2.7') and (2.14') are valid. A workable formula for the standard error of the reliability estimates may be obtained under this model only by employing the well-known characteristic of an F-statistic (Cleary and Linn, 1968). However, if any one or more of the assumptions are not valid the true sample reliability distribution will not be the same as that given by (2.17).

If the ANOVA model equation (2.6') is taken as the basic model, the more general models and the normal ogive model for the binary item test may be considered as being assumption-violating cases of the basic model. It has been shown that the latter more general models are obtained by successively relaxing some of the assumptions of (2.7'), and the normal ogive model has been shown to be the congeneric model if the hypothetical 'response strength' variable is used in the model equation. Thus the problem of investigating the distribution of reliability estimates using models other than the ANOVA model

becomes the problem of investigating the effects of the violation of the assumptions of the ANOVA model upon the distribution of the reliability estimates.

It is also suspected that, under certain circumstances for real data, cases arise in which the distributional assumptions of (2.14') are substantially violated, that is the distribution of the true scores and the error scores may be skewed, and/or platykurtic or leptokurtic (Lord, 1960, 1969).

However, regardless of which model the real data may satisfy, in practice the reliability estimates are usually obtained using the Alpha or KR20 formulas; hence the distributional theory of the estimates based on these formulas becomes a central concern for the test users as well as the theorist. Thus, it seems justifiable to investigate the distributional problem using models other than the ANOVA model, e.g., those in which a systematically distorted distribution arises for (2.17) by violating (2.7') and/or (2.14'). The assumptions underlying such models are summarized in the following table.

TABLE 4.1

Summary of the Assumptions Under Various Models

Assumptions	ANOVA	ETEM	CONG.	M.F.	N.O.[1]
independence of true and error scores	yes	yes	yes	yes	yes
uni-factor true scores	yes	yes	yes	no	yes
ETEM assumptions	yes	yes	no	no	no
homogeneity of error score variances	yes	no	no	no	no
normality of true scores	yes	yes	yes	yes	yes
normality of error scores	yes	yes	yes	yes	yes

(Cong. - congeneric; M.F. - multi-factor; N.O. - normal ogive)

[1]Applicable only to the response strength variable.

4.2 Robustness Under Violation of Assumptions

It has been known that, under certain conditions, the F-test of the one way fixed effects analysis of variance model is quite robust against the violation of the underlying assumptions. It may then be asked whether or not the same robustness exists for inferences about the reliability based on (2.17), which relies on an F-statistic. That is, can the findings for the one way fixed

effects model ANOVA be generalized to the two way mixed effects model ANOVA case with one observation per cell. If the sampling distribution of the reliability estimates is stable with the violation of assumptions, statistical inference based on (2.17) would be very powerful. If the sampling distribution of the most often used Alpha or KR20 estimates are found to be quite robust against the violation of the assumption users may freely employ the Alpha and KR20 estimate formulas and perform statistical inferences based on (2.17) without investigating the adequacy of the models or the assumptions. If the distribution is robust only under certain conditions, the researcher should keep this in mind whenever making an inference about the reliability or interpreting an estimate based on (2.17). Therefore, the basic question to be answered is: under what conditions, if any, do the Alpha or KR20 estimates have a stable distribution against the violation of assumptions.

4.3 An Empirical Approach Toward the Problem

Since a mathematical answer to the above problem is not available at the moment, and it seems impossible to give one in the near future, one alternative approach to be considered is the performance of an actual experiment, i.e., an empirical examination of the sampling distribution under various models and assumptions that violate the ANOVA model and its assumptions. The empirical distribution of the Alpha and KR20 estimates can then be found and compared with the theoretical one under the ideal ANOVA model.

An experiment with real data is almost impossible since the population parameters are seldom known. Even if this were possible

the data would not fit the specific model and assumptions except for rather limited cases (e.g., Baker, 1962). One available method is to use computer simulated data, under various assumptions, to obtain empirical distributions of the Alpha and KR20 estimates and compare them with the distribution for the ideal ANOVA model.

The author has already investigated the feasibility of such computer simulation techniques in the study of the effects of the violation of assumptions on the F-test for linear models requiring statistical inferences, and has provided a comprehensive computer program for educational and psychological researchers (Bay, 1970).

The present study uses essentially the same techniques to investigate the sampling distribution of reliability estimates.

4.4 The Concept of Simulation

The term 'simulation' has been used rather uncritically in a wide range of scientific or economic fields, especially for the purpose of building models. Von Neuman and Ulam's work in the late 1940's, when they attempted to solve certain nuclear physics problems by a Monte Carlo analysis, may be considered the first modern use of the simulation techniques. A Monte Carlo analysis involves the solution of a problem, that is either too expensive for experimental solution or too complicated for analytical methods, by simulating a stochastic process that has probability distributions satisfying the mathematical or probabilistic relations underlying the problems.

With the development of high speed computers in the last two decades, not only physicists and other natural scientists, but also

economists, psychologists, and other social scientists can perform controlled laboratory-like experiments on a computer with much efficiency and economy.

Although there is no agreed upon definition of the term 'simulation', for the purpose of this paper it was considered sufficient to use the following definition given by Churchman (1963, p. 12).

> 'x simulates y' is true if and only if:
>
> (a) x and y are formal systems,
>
> (b) y is taken to be the real system,
>
> (c) x is taken to be an approximation to the real system, and
>
> (d) the rules of validity in x are non-error free.

In the context of this paper, y is a system which produces a number of real test score sets by performing actual random sampling of subjects and administering the test, thus giving a number of real estimates of reliability of the test. The number of estimates under the real situation is limited since the actual sampling of subjects and the administration of the test are involved. On the other hand, x is a system which produces a number of test score sets, via computer, under a model and a number of assumptions which will approximate the real system y. Since the number of test score sets obtainable under x is almost unlimited, the sampling distribution of the relaibility estimates is easily obtainable by calculating the frequency distribution of the estimates. Furthermore, since the test parameters and the distributions of true and error scores can be manipulated easily under computer simulation, almost any combination of models and assumptions can be investigated. The researcher can input the most appropriate model

and assumptions which will best approximate the real system y for a given test and population of subjects.

This approach toward statistical inference is somewhat different from the conventional procedure since the user can choose the model and assumptions of interest to him, while in the conventional case the model and assumptions are predetermined by the mathematical statisticians and the user can only choose whether or not to accept the conditions and the model, look for alternatives, or give up. In this sense computer simulation techniques permit the study of sampling distributions under almost unlimited combinations of models and assumptions. Thus, the user may obtain the sampling distribution of a statistic under his own model and assumptions in the experimental situation, make statistical inferences, and use the knowledge so gained in practice. Because of the fourth property of the simulation, the method may not provide exact answers, but it would provide approximate answers to distribution problems.

4.5 Computer Programs

Two computer programs named REL01 and REL02 have been developed in FORTRAN IV on the IBM 360/67 computer of the University of Alberta computer system for continuous part test and binary item test cases respectively. The programs are in sufficient general form so that they can be used for other problems related to sampling distribution of reliability estimates not considered part of the study. The programs automatically simulate the test score matrix $\underline{Y} = \{y_{ij}\}$ for the continuous case, or $\underline{X} = \{x_{ij}\}$ for the binary case based on input models, parameters, and specified distributions of true or latent

scores and error scores. The programs have the following features:

(a) For the continuous case, the program REL01 uses the most general model, namely the multi-factor true score model given by (2.18), and accommodates all other less general models as special cases. Users are able to specify the sample size I, the number of parts J, and the parameter vector and matrices for the model, namely $\underline{\mu}$, $\underline{\Lambda}$, and $\underline{\Psi}$, i.e., mean vector, factor loading matrix, and error standard deviation matrix respectively. The program will evaluate population test parameters such as reliability, Alpha, mean, true and error variances.

(b) For the binary item case, the program REL02 uses the normal ogive model (3.1), under a uni-factor latent scores assumption, and allows the user to specify the sample size I, the number of items J, and the basic item parameters, namely the difficulty parameters $\{\pi_j\}$ and the biserial correlations $\{\lambda_j\}$. The program will evaluate the population test parameters such as ρ, KR20, and σ_x^2 based on the Tchebycheff-Hermite polynomials discussed in Chapter Three and other formulas under the normal ogive model. However, these calculations are valid only for the normal distributions of latent scores and errors. If the normality is violated, the parameters calculated are no longer valid, unlike the case of continuous parts where the test parameters are independent of distributions of true and error scores. To evaluate test parameters for non-normal cases, an empirical method based on a parallel form method described in the following Section 4.6 is used.

(c) For both programs the user may decide shapes of the

distributions of true or latent scores $\{f_i\}$ and error scores $\{\varepsilon_{ij}\}$. The programs generate specific distributions by means of random number generating subroutines. The distributions of true or latent scores and error scores are specified by user supplied subroutines DIST and DISE respectively for non-normal cases. These two subroutines may call the uniform random number generating subroutine VECRAN described in the following Section 4.7. For the normal case, the program generates the distribution automatically by employing the Box-Muller method which is also described in the Section 4.7.

(d) The programs automatically perform N simulations, as specified by the user, and calculate a number of test statistics. The reliability coefficients are estimated for each simulated test score matrix based on the formula (2.13), regardless of the model and distributions used to generate the score matrix, since the formula is, as was noted before, the one most often used by the test theorists or users. Alternatively or concurrently, as an option, the user may adopt an unbiased estimation formula developed by Kristoff (1963) and discussed in the following chapter. The distributions of reliability thus estimated are then compared with those obtainable from (2.17), i.e., the ideal ANOVA model and normal theory. For non-normal binary item test cases, the reliability parameter obtained by the parallel form method is used for the value of ρ.

(e) The programs also summarize the empirical distributions of MS_A, MS_B, MS_e, and $\hat{\rho}$ by calculating their means and variances over N samples, and compares them with the theortical values of the expected mean and variance under the ANOVA model and normal theory

assumptions. For the binary item case, the variance parameter σ_A^2 and σ_e^2 in terms of the test score x_i are not defined or calculable directly. However, from the definition of reliability given by (2.12) and the variance given by (2.10), a generalization of the relationships between reliability and variances to binary item cases may be made such that the formulas in Chapter Two may be used without modification, namely,

$$(4.1) \qquad \sigma_A^2 = \frac{\rho\sigma_x^2}{J^2} \; ; \qquad \sigma_e^2 = \frac{(1-\rho)\,\sigma_x^2}{J} \; , \quad \text{or}$$

$$(4.1') \qquad \sigma_A^2 = \frac{\rho^* \sigma_x^{*2}}{J^2} \; , \qquad \sigma_e^2 = \frac{(1-\rho^*)\sigma_x^{*2}}{J} \; ,$$

for non-normal cases, where the star (*) notation referes to parameters evaluated by the parallel form method.

(f) Comparisons between the empirical distributions of reliability estimates based on either or both (2.13) or Kristof's unbiased formula, with those theoretical ones based on (2.17) or modified form of it for the unbiased formula, can also be made as an option by plotting both distributional curves together in a graph.

Computer program listings together with example outputs of the programs are given in Appendix A.1 and A.2 respectively.

4.6 Parallel Forms Method for Test Parameters of Binary Item Test

For the continuous part test cases, test parameters such as σ_y^2, ρ, and Alpha depend only on the input of part test parameters and are independent of the distributions of the true and error scores.

However, for the binary item test cases, the basic test parameters depend not only on item parameters such as difficulty or biserial correlations but, also on the distributions of latent scores and errors, since the normal ogive model connects the continuous response strength variable y_{ij} to the binary item score x_{ij}. Therefore the formulas for test parameters such as σ_x^2, ρ, and KR20 given in Chapter Three are valid only for the case of normal distributions. In order to be able to investigate the sampling distributions of reliability estimates under non-normal cases, i.e., under the assumption violating cases of the normal ogive model, the test parameters must be known by means other than these formulas. Although for some simpler distributions such as the uniform distribution, evaluation of these parameters by analytical means might be possible, a general solution to cover all types of possible distributions is impossible, and alternative empirical methods are employed in the REL02 program.

Since the number (N) of test score matrices simulated is usually large, say at least 1000, the number of test scores $(N \times I)$ simulated in each experiment is a very large number compared with the sample size I. On the other hand, the sample reliability and variance are consistent estimators of the corresponding population values. Therefore, if $N \times I$ test score sets are used at a time to estimate these parameters, the estimates will be close to the population values. However, to obtain a population reliability coefficient by this large sample method and the correlation formula give by (3.12), parallel form test scores must also be simulated which have identical f_i terms but different ε_{ij} terms denoted by ε_{ij}^* due to random

fluctuation of responses. Therefore, two sets of model equations may be considered for the response strength variables y_{ij}, namely,

$$(4.2)\qquad \begin{cases} y_{ij} = \lambda_j f_i + (1 - \lambda_j^2)^{\frac{1}{2}} \varepsilon_{ij}, \\ \\ y^*_{ij} = \lambda_j f_i + (1 - \lambda_j^2)^{\frac{1}{2}} \varepsilon^*_{ij}; \qquad i = 1,\ldots,NI; \quad j = 1,\ldots,J. \end{cases}$$

Form these model equations, two sets of test scores $\{x_i\}$ and $\{x^*_i\}$ may be generated, and by calculating the correlation coefficient between these two sets of scores, the population reliability may be obtained regardless of which distribution is used for simulating the test scores. For ideal normal cases, the parameters obtained by this method should agree closely with the calculated values based on the formulas of Chapter Three, providing one way of checking the formulas in the chapter and the computing procedures adopted by REL02. The population parameters thus estimated will be denoted by the corresponding population parameter symbols with a star (*) sign to distinguish them from those obtained by analytical means. For example, the test mean and variance obtained by this method are given by,

$$\mu^* = (\textstyle\sum_i x_i)/NI, \qquad \text{and} \qquad \sigma_x^{*2} = \{\textstyle\sum_i (x_i - \mu^*)^2\}/NI \ .$$

4.7 Procedures for Generating Random Numbers

The method of generating a set of independent random numbers with a specific distribution by a computer program is of extreme importance to the success of a stochastic simulation experiment. The

simplest and basic set of random numbers with a continuous probability density function is the one that is constant over the interval (0,1) and is zero otherwise. The density function defines what is known as a uniform or square distribution. The principal value of the uniform distribution for the simulation techniques lies in its simplicity and in the fact that it can be used to simulate random variables from almost any kind of probability, distribution since the inverse transformation of the cummulative distribution function of any random variables results in the uniform distribution between (0,1).

The uniform density function on (0,1) is defined by

$$(4.3) \qquad f(z) = 1.0 \qquad 0 < z < 1$$
$$= 0.0 \qquad \text{otherwise.}$$

Due to its simple density function, it is very easy to evaluate moments for such a uniformly distributed random variable by using elementary calculus.

For this study, the method used for generating uniform random number is the same as that used by the IBM Scientific Subroutine Package RANDU (IBM, 1968). The subroutine named VECRAN can however generate a specified number of uniform random numbers at a time and provides the output in vector form, while only one number at a time is generated by RANDU.

The method employed is the so-called 'power residue method', (IBM, 1959) or 'multiplicative congruential method' (Nayler et al., 1968, p. 51-52). The method generates successive non-negative integer random number which are less than 2^c for binary computers where c denotes the word size of the computer by means

of a congruence relation, namely,

$$n_{i+1} = a\, n_i \text{ (modulo of } 2^c\text{)}, \quad i = 0,1,\ldots \tag{4.4}$$

where n_0 is the so-called seed random number denoted by IX in the program. Meanings of 'power residue', 'congruential' or 'modulo' are given by Nayler et al., (1968, pp. 63-66), or can be found in any text-book of elementary number theory. The formula (4.4) is the so-called formula for generating power residuals, and results in $u = n_{i+1}/(2^c-1)$ being approximately a uniform random number in $(0,1)$. For the IBM 360 series computers, $c = 31$, and VECRAN uses $a = 65539$, and $2^{-c} = 0.4656613 \times 10^{-9}$ which are the same as for RANDU. The user must specify $n_0 = \text{IX}$ as an input parameter at the beginning of the program execution, and it must be an odd integer with nine digits or less. The last value of n_i generated may be used as an input seed random number IX for the next step generation.

The random numbers thus generated are often referred as pseudo-random numbers, and the method involves the generating procedure by 'indefinitely continued transformation of a group of arbitrarily chosen numbers' (Tocher, 1954, p. 41). The term has been defined by Lehmer (1951) as,

> ... a vague notion embodying the idea of a sequence in which each term is unpredictable to the uninitiated and whose digits pass a certain number of tests, traditional with statisticians and depending somewhat on the use to which the sequence is to be put.

Although there are some objections on the philosophical grounds that the sequence is generated by a deterministic rule of (4.4), use of such pseudo-random numbers can be defended by pragmatic reason that a sequence may be regarded random if it satisfies some predetermined

statistical tests of randomness, and the uniform number generated by RANDU has been known to satisfy these requirements (IBM, 1968).

Based on the uniform random numbers thus generated by VECRAN, denoted by U1, five other kinds of random numbers are generated for this study. For the selection of these specific types of a random number the following factors were taken into account:

(a) Ease of generation and computer time required for computation.

(b) Ease of evaluating the moments of random numbers by calculus to ensure that the program generates random numbers with the required distribution.

(c) Some practical usefulness. For example, normal, uniform, and exponential distributions are included because the approximation of the normal distribution to real data is so often assumed, the uniform distribution is closely associated with ranked data, and the exponential distribution can arise with the truncated data of normal distribution due to a selection process.

The six kinds of random numbers, including U1, are summarized in the following table.

TABLE 4.2

Summary of the Random Numbers

Description	Notation	Transformation Formula
Uniform, (0,1)	U1	$z = u_1$
Sum of 2 indep. U1	U2	$z = \{(u_1+u_2) - 1.0\} \times (6)^{\frac{1}{2}}$
Sum of 3 indep. U1	U3	$z = \{(u_1+u_2+u_3) - 1.5\} \times 2$
Sum of 6 indep. U1	U6	$z = \{(u_1+u_2+\ldots+u_6) - 3.0\} \times (2)^{\frac{1}{2}}$
Normal	NO	$z_1 = (-2 \text{ Ln } u_1)^{\frac{1}{2}} \text{ Cos } (2\pi u_2)$ $z_2 = (-2 \text{ Ln } u_1)^{\frac{1}{2}} \text{ Sin } (2\pi u_2)$
Exponential	EX	$z = -\text{Ln } (u_1) - 1.0$

Note: $u_1,\ldots,u_6$ denote the uniform random numbers generated by VECRAN.

The method used for the generation of standard normal random variables is the same as given by Box and Muller (1959). Since the distribution is exact, it has an advantage over the so-called central limit approach which uses the sum of a number of independent uniform random variables. All random variables in this study were used in standard form, namely with an expected value of zero and unit variance, except U1 which was standardized by subtracting 0.5 and multiplying by the square root of 12.0. Therefore the random numbers thus generated can easily be used as $\{f_i\}$ or $\{\varepsilon_{ij}\}$ of the model equations (2.28), (3.3), and (4.2).

A number of preliminary sampling experiments were performed to ensure that this method generates the random numbers with desired distributions. To see whether the means, variances and other statistics for large samples closely approximated the population values of the distribution simulated by the random numbers, five samples of size 6000 each were generated for each distribution, and the obtained statistics were compared with the population values obtained from the knowledge of the probability density functions and the application of elementary calculus. The results are summarized in Table 4.3. With some exceptions for the calculated kurtosis of the distribution noted with (*) sign, the sample statistics approximate reasonably well the population values. The exceptional cases are probably due to the imperfections of the random number generating procedures and sensitivity of kurtosis to the shapes of the distribution. The calculated auto-correlations are almost zero indicating no serial correlations for adjacent random numbers in the sequences and the degree of independence of random numbers thus generated.

4.8 Methodological Limitations

Because the computer simulated experiments cannot be exhaustive and cover all possible combinations of models, parameter sets, and distributional assumptions, and due to the very nature of computer simulation techniques and limited funds available for the computing charges, the following methodological limitations are imposed on this study.

TABLE 4.3

Descriptive Summary of Random Numbers Generated by Pseudo-Random Number Generating Subroutines, Sample Size = 6000 for Each Trial

Dis	Trial	Mean	Var.	Skewness	Kurtosis	Auto-Correlations		
						Lag 1	Lag 2	Lag 3
U1	1	0.50347	0.08417	-0.01615	-1.21110	0.00153	-0.00182	0.00152
U1	2	0.50027	0.08500	-0.00620	-1.23131	0.00029	0.00054	-0.00206
U1	3	0.50080	0.08367	-0.02638	-1.20730	-0.00070	0.00045	-0.00049
U1	4	0.49840	0.08320	-0.01426	-1.19213	0.00036	0.00174	0.00073
U1	5	0.50682	0.08226	-0.02995	-1.18213	0.00118	0.00052	0.00005
U1	Expected	0.50000	0.08333	0.00000	-1.20000	0.00000	0.00000	0.00000
U2	1	0.03035	0.99873	-0.00577	-0.58865	0.02595	-0.22086	-0.00478
U2	2	0.01817	0.98648	-0.02644	-0.59803	0.01458	-0.00631	0.00929
U2	3	-0.00268	1.01429	0.00608	-0.63365	0.00781	0.02034	-0.00341
U2	4	0.01204	0.98045	-0.02414	-0.58109	-0.01133	0.01728	0.00093
U2	5	0.02277	1.01143	-0.00802	-0.65140	0.02078	0.02055	0.00163
U2	Expected	0.00000	1.00000	0.00000	-0.60000	0.00000	0.00000	0.00000
U3	1	0.02933	1.00741	0.00310	-0.42060	0.02305	-0.01435	0.01603
U3	2	0.01852	0.95938	0.00809	-0.30457	0.01778	-0.01994	0.00278
U3	3	-0.00475	1.00159	-0.01646	-0.37746	0.00498	0.01504	-0.00118
U3	4	0.00062	0.99190	-0.02163	-0.37966	-0.00196	0.00308	-0.00630
U3	5	0.01843	1.02359	-0.01513	-0.40427	0.00996	0.00901	-0.00601
U3	Expected	0.00000	1.00000	0.00000	-0.40000	0.00000	0.00000	0.00000
U6	1	0.01903	0.98999	-0.01882	-0.23413	-0.01181	-0.01000	0.02742
U6	2	0.00843	0.97866	0.00333	-0.20790	-0.00566	-0.02743	0.02315
U6	3	-0.00411	1.00637	-0.00300	-0.22061	-0.01835	-0.01835	0.02315
U6	4	0.00294	0.99037	-0.01588	-0.24665	-0.00398	0.00064	-0.01555
U6	5	0.01480	1.03165	-0.00895	-0.16114	0.00049	0.01274	0.00500
U6	Expected	0.00000	1.00000	0.00000	-0.20000	0.00000	0.00000	0.00000
NO	1	0.00587	1.01532	-0.00714	0.11501*	0.01517	0.01176	-0.00349
NO	2	0.00214	0.99160	-0.05184	0.05017	-0.00904	0.02503	-0.01246
NO	3	-0.00856	0.99918	-0.00635	0.03248	0.02173	-0.02119	0.03329
NO	4	-0.00069	0.99955	0.02294	-0.04551	0.01060	-0.02274	-0.00699
NO	5	-0.01763	0.97289	0.01456	-0.05775	-0.01532	-0.00687	0.00793
NO	Expected	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
EX	1	-0.00475	1.01587	2.00600	5.70110	0.00894	-0.01838	0.01769
EX	2	0.00565	1.01611	1.93136	5.31563	0.00892	0.01781	-0.01341
EX	3	0.00586	1.05517	2.19383	8.02954*	-0.00005	-0.00905	-0.01218
EX	4	0.00866	1.01780	1.91303	4.91614*	0.00727	0.01803	-0.00044
EX	5	-0.02518	0.96299	2.03853	6.10958	0.00473	-0.00114	0.00189
EX	Expected	0.00000	1.00000	2.00000	6.00000	0.00000	0.00000	0.00000

Note: All random numbers are standardized by population mean and variance except U1.

(a) The investigation is restricted to the sampling distributions of reliability estimates under Type 1 sampling situation only, namely only sampling of subjects is involved; the test is assumed to be given and all parameters for part-tests or items are assumed to be fixed constants. The distributions under the Type 2 or Type 12 sampling situation, such as the distributions of generalizability coefficient estimates, are not considered in this study, although this may be done very easily as an extension to this study.

(b) Because the computer time required for each experiment must be kept within reasonable limits, the sample size I, the number of parts or items J, and the number of samples to be simulated must be kept within moderate bounds for this study. Therefore, although the programs are dimensioned such that they can accommodate up to $N = 5000$, $I = 100$, $J = 30$, investigations are limited to $N = 2000$ or 1000, $I = 30$, $J = 6$, 8, or 9 to restrict each experiment within 5 to 7 minutes of C.P.U. time which costs approximately $20-30 at the present charging rate of the University of Alberta.

(c) To conserve computer time for the overall study, the experiments have focused only on the following key problems:

(i) The effect of non-normal true or latent scores and error scores distributions.

(ii) The effects of non-homogeneous error variances, i.e., distributions under ETEM model.

(iii) The effect of congeneric and multi-factor true score model.

(iv) The effect of binary item scores, non-homogeneous difficulty parameters and biserial correlations, and non-normal distributions.

(d) The non-normal distributions used in this study are limited to a minimal number of well known distributions outlined in Section 4.7.

Because of these limitations, the findings of this study will be limited to some extent in their generalization to all 'real' situations.

4.9 Accuracy of Calculation

Like any other numerical analysis, the results reported in this study are subject to certain computational errors. The figures reported in this study retain, in most cases, three decimal places, but they may be inaccurate in the right most one significant digit due to the cummulative effects of errors when the sample size N is large. This is especially true for the case when the variance of a variable is small in comparison with the mean. However, it is expected that the errors are confined only within 3 to 4% level at maximum, and they would not affect the findings of this study.

4.10 Summary

In this chapter, the rationale for investigating the sampling distributions of reliability estimates as assumption violating cases of the well known ANOVA model and normal distributional theory, and using the computer simulation technique to investigate such

problems were discussed. The computer programs developed for this study were outlined, and the parallel form method, the random number generating procedures, and the methodological limitations due to the very nature of computer simulation techniques were also discussed.

CHAPTER FIVE

RESULTS FOR CONTINUOUS PART TEST SCORE CASES

This chapter presents the results of the computer simulated experiments for the continuous part test score cases. Section 5.1.0 deals with the effects of non-normality under the ANOVA model; and some analytical methods are also used to investigate the standard error of reliability estimates. The distributions of reliability estimates under the ETEM model are dealt with in Section 5.2.0, and in Section 5.3.0 for the congeneric and multi-factor true score cases, i.e., non-ETEM cases.

5.1.0 Effects of Non-Normality Under the ANOVA Model

5.1.1 Distribution Under ANOVA and Normal Distribution of True and Error Scores

It has been shown in Chapter Two that, under the ANOVA model and normal distribution, the reliability estimate given by (2.13)-(b) can be related to an F-statistic by the equation (2.17), and it can also be shown that (Kendall and Stuart, 1963, p. 393),

$$E(F_{m;n}) = \frac{n}{n-2}\,, \qquad Var\ (F_{m;n}) = \frac{2\,n^2(n+m-2)}{m(n-2)^2(n-4)}\,.$$

Therefore, using the relation $1/F_{m;n} = F_{n;m}$, it is easy to show that,

$$(5.1)\quad \begin{cases} \text{(a)}\quad E(\hat{\rho}) = 1 - (1-\rho)\ E(F_{\nu;I-1}) = 1 - (1-\rho)\ \dfrac{I-1}{I-3} \\[2ex] \text{(b)}\quad Var\ (\hat{\rho}) = (1-\rho)^2\ \dfrac{2(I-1)(\nu+I-3)}{(J-1)(I-3)^2(I-5)}\,. \end{cases}$$

Hence $\hat{\rho}$ is in general a biased, but consistent estimator and does not have the minimum variance property. Kristof (1963) modified formula (2.13) to obtain the unbiased estimator $\hat{\hat{\rho}}$ and has shown that it has a smaller variance than $\hat{\rho}$, namely

$$(5.2)\quad \begin{cases} \text{(a)}\quad \hat{\hat{\rho}} = \dfrac{2}{I-1} + \dfrac{I-3}{I-1}\,\hat{\rho} = \dfrac{2}{I-1} + \dfrac{I-3}{I-1}\,(1 - MS_e/MS_A)\ , \quad \text{or} \\[2ex] \text{(b)}\quad F_{I-1;\nu} \doteq \dfrac{(I-3)(1-\rho)}{(I-1)(1-\hat{\hat{\rho}})} \quad . \end{cases}$$

It can then be easily shown that

$$E(\hat{\hat{\rho}}) = \rho; \qquad \text{Var}\ (\hat{\hat{\rho}}) = [\frac{I-3}{I-1}]^2\ \text{Var}\ (\hat{\rho}) = (1-\rho)^2\ \frac{2(\nu+I-3)}{\nu(I-5)} \leq \text{Var}\ (\hat{\rho})\ .$$

Therefore, if the equation (2.6) is the appropriate model for the data and the assumptions (2.7) and (2.14) are all satisfied, the results of equations (2.17), (5.1), or (5.2) can be used to make inferences about ρ and to calculate the standard error of estimation which is defined as the square root of the variance of $\hat{\rho}$.

5.1.2 Known Effects of Non-Normality Under ANOVA

As it has been seen, the sampling theory and the formula for the standard error of estimation rely heavily on the normal distribution assumptions, despite the fact that real data seldom satisfy these assumptions, and at best may be expected to only approximately satisfy them. It does not logically follow, of course, that approximate satisfaction of the normal distribution assumptions by true and error scores will guarantee automatic approximation of the actual distribution of reliability estimates to the distribution given under normal theory.

Scheffé (1959, p. 345) investigated the effect of non-normality from an analytical point of view and concluded 'Non-normality has little effect on inferences about means but serious effects on inferences about variances of random effects whose kurtosis γ_2 differs from zero'. He also noted that 'The direction of the effect is such that for confidence coefficients $1-\alpha$ and significance level α the true α will be less than the nominal α if the $\gamma_{2,A} < 0$, and greater if $\gamma_{2,A} > 0$, and the magnitude of the effect increases with the magnitude of $\gamma_{2,A}$.' Although his argument is based on the inference of the so-called signal-noise ratio $\theta = \sigma_A^2/\sigma_e^2$, under the one way random effects model, it is suggestive for reliability theory, and provides a guideline for the investigation of the effects of non-normality under the ANOVA model.

5.1.3 Standard Error of Reliability Estimates Corrected for Non-Normality

The standard error of reliability estimates is a useful measure of the precision of the estimates, although, as noted in Chapter One, without any knowledge of the shape of the sampling distributions of the estimates it has little inferential use. Since reliability has been historically identified as a correlation coefficient, the well-known standard error of correlation coefficient estimates has been frequently used (e.g., Jackson and Ferguson, 1941), namely,

$$\text{(5.3)} \qquad \text{Var}\ (\hat{\rho}) = \frac{(1-\rho^2)^2}{I}$$

in which the assumption of bivariate normality is made (Kendall and Stuart, 1963, p. 236). However, this formula or those given by equations (5.1) and (5.2) would be misleading if normality cannot be

assumed. General distributional theory under non-normal true and error scores is not yet known, but the Var ($\hat{\rho}$) or its square root, denoted by S.E. ($\hat{\rho}$), may be evaluated approximately if the kurtosis of the true and error scores, denoted by γ_A and γ_e respectively, are known or can be estimated. In this case

$$(5.4)\quad \begin{cases} \text{(a)} \quad \gamma_A = [E(a_i^4)/\sigma_A^4] - 3, \\ \\ \text{(b)} \quad \gamma_e = [E(e_{ij}^4)/\sigma_e^4] - 3. \end{cases}$$

Tukey (1956) obtained the variance of the variance estimates under various ANOVA models by employing 'polykays'. For the model considered in this paper, he has shown that

$$(5.5)\quad \begin{cases} \text{(a)} \quad \mathrm{Var}\ (\hat{\sigma}_A^2) = \frac{2}{I-1}\,\sigma_A^4 + \frac{4}{J(I-1)}\,\sigma_A^2\,\sigma_e^2 + \frac{2}{J(J-1)(I-1)}\sigma_e^4 + \frac{\gamma_A}{I}\,\sigma_A^4\ , \\ \\ \text{(b)} \quad \mathrm{Var}\ (\hat{\sigma}_e^2) = \frac{2}{(I-1)(J-1)}\,\sigma_e^4 + \frac{\gamma_e}{IJ}\,\sigma_e^4\ , \\ \\ \text{(c)} \quad \mathrm{Cov}\ (\hat{\sigma}_A^2,\ \hat{\sigma}_e^2) = \frac{-2}{(I-1)(J-1)J}\,\sigma_e^4\ . \end{cases}$$

From (5.5) it is easy to obtain

$$(5.6)\quad \begin{cases} \text{(a)} \quad \mathrm{Var}\ (MS_A) = [\frac{2}{I-1} + \frac{1}{I}\{\rho^2\,\gamma_A + (1-\rho)^2\,\gamma_e/J\}](J\,\sigma_A^2 + \sigma_e^2)^2 \\ \\ \text{(b)} \quad \mathrm{Var}\ (MS_e) = [\frac{2}{(I-1)(J-1)} + \frac{\gamma_e}{IJ}]\,\sigma_e^4 \\ \\ \text{(c)} \quad \mathrm{Cov}\ (MS_A,\ MS_e) = \frac{\gamma_e}{IJ}\,\sigma_e^4\ . \end{cases}$$

It is noted that if the true and error scores are normal, i.e., the kurtosis is equal to zero, the results are the same as expected

under normal theory obtainable from equation (2.15) and the resulting independence of MS_A and MS_e.

Letting $x_1 = MS_A$, and $x_2 = MS_e$, and $W(x_1,x_2)$ a function of x_1 and x_2, an approximation formula (e.g., Scheffé, 1959, p. 230) may be applied to approximate Var $(\hat{\rho})$ from (2.13)-(b) namely,

$$\text{Var}\ [W(x_1,x_2)] \simeq W_1^2\ \text{Var}\ (x_1) + 2W_1W_2\ \text{Cov}\ (x_1,x_2) + W_2^2\ \text{Var}\ (x_2)\ ,$$

where W_i denotes $\partial W/\partial x_i$ evaluated at $x_1 = E(x_1) = J\ \sigma_A^2 + \sigma_e^2$, and $x_2 = E(x_2) = \sigma_e^2$. Then, Var $(\hat{\rho})$ = Var $(1 - MS_e/MS_A)$ = Var (x_2/x_1), i.e., $W(x_1,x_2) = x_2/x_1$, and $W_1 = -\sigma_e^2/(J\ \sigma_A^2 + \sigma_e^2)^2$, $W_2 = 1/(J\sigma_A^2 + \sigma_e^2)$, giving

$$(5.7) \qquad \text{Var}\ (\hat{\rho}) \simeq (1-\rho)^2\ [\frac{2J}{(I-1)(J-1)} + \frac{\rho^2}{I}\ (\gamma_A + \gamma_e/J)]\ .$$

Formula (5.7) does not agree exactly with formula (5.1)-(b) when the distributions are normal since an approximation has been employed. However, formula (5.7) is suggestive for correction terms to be added to formula (5.1)-(b) for non-normal distributions, i.e., Var $(\hat{\rho})$ may be obtained by a new formula combining (5.1) and (5.7) as

$$(5.8) \qquad \text{Var}\ (\hat{\rho}) \simeq (1-\rho)^2\ [\frac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \frac{\rho^2}{I}\ (\gamma_A + \gamma_e/J)]\ .$$

Since this formula involves two unestimable parameters γ_A and γ_e, further approximation is necessary to make it useful.

The kurtosis of the test scores $y_i = \sum_j y_{ij} = J\mu + J\ a_i + \sum_j e_{ij}$, denoted by γ_y, is an estimable parameter, and may be evaluated by considering it as a linear combination of $J+1$ independent

random variables a_i and $\{e_{ij}\}$ for $j = 1,\ldots,J$, and applying a formula given by Scheffé (1959, p. 332), namely,

$$(5.9) \qquad \gamma_y = \rho^2 \gamma_A + (1-\rho)^2 \gamma_e/J \quad .$$

Then, $\gamma_y \simeq \rho^2 \gamma_A$ for $\rho \simeq 1$, or $\gamma_e \simeq 0$, or J fairly large. Therefore, it may be shown that,

$$(5.10) \qquad \mathrm{Var}\ (\hat{\rho}) \simeq (1-\rho)^2 \left[\frac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \frac{\gamma_y}{I}\right] .$$

This formula (5.10) is, to the author's knowledge, a new one for test theory, which only includes the known constants I,J and the unknown but estimable parameters ρ and γ_y. As a result it can be used to obtain an estimate of the standard error of reliability estimates, namely,

$$(5.11) \qquad \widehat{\mathrm{S.E.}\ (\hat{\rho})} = [\widehat{\mathrm{Var}\ (\hat{\rho})}]^{\frac{1}{2}} \simeq (1-\hat{\rho}) \left[\frac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \hat{\gamma}_y/I\right]^{\frac{1}{2}} .$$

From the formula (5.8) it may be observed that the effects of non-normality on the standard error of reliability estimates depend on the following:

(a) The kurtosis of the true scores multiplied by the factor $1/I$, and of the error scores multiplied by a factor of $1/IJ$. Therefore, the effect of non-normality would be dominated by the kurtosis of true scores which is closely approximated by the kurtosis of the test scores divided by the square of the reliability.

(b) The magnitude of ρ, namely, the larger the value of ρ, the greater is the effect of non-normality.

The above observations suggest that the sampling distribution would be robust against the violation of normality assumptions if (a) the sample size I is large, (b) reliability is close to zero, or (c) J is fairly large and the true score kurtosis (or the test score kurtosis) is close to zero. The condition (a) is of little practical value since statistical inference problems usually arise for the small sample case, while (b) is also of little practical value since, in most cases, reliability theory deals with ρ close to unity rather than zero. The last condition indicates that the sampling distribution of reliability estimates would be robust against the violation of normality of errors for J fairly large, and is sensitive to the distribution of true scores.

5.1.4 Results of Simulation Experiments Under ANOVA Model

In order to investigate the effect of non-normality under the ANOVA model, a number of experiments were performed by REL01 using the following distribution-parameters combinations with the constants $N = 2000$, $I = 30$, and $J = 8$.

(a) For the distribution of true scores, all of the six distributions discussed in Table 4.2 of Chapter Four, namely U1, U2, U3, U6, NO, and EX. were used.

(b) For the error scores distributions, the uniform, normal, and exponential distributions were used, i.e., U1, NO, and EX respectively.

(c) Three levels of ρ were used by fixing $\sigma_e^2 = 4.0$, and using three levels of σ_A^2, namely, 4.0, 1.0, and 0.36 to indicate high, middle and lower levels of reliability.

Altogether $6 \times 3 \times 3 = 54$ experiments were performed, each

requiring approximately six minutes of C.P.U. time. Since the parameters μ and $\{\beta_j\}$ do not affect the distributions, they are not reported.

In Table 5.1, the observed means and variances of MS_A and MS_e for $N = 2000$ samples are presented with the theoretical values based on formula (5.6). Because formula (5.6) does not involve any approximation, any disagreement between the observed and calculated values must be attributed to either sampling fluctuations due to the finiteness of N or deficiencies in random number generating methods. It is noted that a rather close agreement exists between the observed means of MS_A and MS_e given in columns (1) and (3) with their theoretical expected values given in column (7). Comparisons of the observed variances of the MS's given in columns (2) and (4) with the theoretical values based on (5.6) given in columns (5) and (6) suggest that the two agree reasonably well, although the agreement is not as close as that for the means and expected values, which probably reflects the imperfectness of the random number generating procedures and/or the sensitivity of the variance to the change in the shape of population distributions.

Column (1) of Table 5.2 contains the mean of $\hat{\rho}$ over the N samples. These values can be compared with the expected values under normal distribution theory given in column (6). It is observed that, for negative γ_A, the means are in general higher than $E(\hat{\rho})$ based on formula (5.1)-(a), thus causing some moderating in the tendency to underestimate the reliability under normal thoery. If γ_A is positive, the mean of $\hat{\rho}$ is in general lower than $E(\hat{\rho})$ and exaggerates the tendency of underestimation. The effect of γ_e is

[illegible]

TABLE 5.1

Comparisons of Observed Means and Variances of MS's Under ANOVA Model and Various Combinations of True and Error Score Distributions With the Values Obtainable From Formula (5.6), N = 2000, I = 30, J = 8

Ex. No.	Dis. Tr.	Er.	Observed MS_A Mean (1)	Observed MS_A Var. (2)	Observed MS_e Mean (3)	Observed MS_e Var. (4)	Var. by (5.6) MS_A (5)	Var. by (5.6) MS_e (6)	Parameters and E(MS) (7)
01	U1	U1	36.052	47.108	4.005	0.085	48.339	0.078	
02	U1	NO	36.102	45.563	3.998	0.162	48.419	0.158	$\sigma_A^2 = 4.0$
03	U1	Ex	36.057	45.041	3.972	0.523	48.819	0.558	
04	U2	U1	36.286	66.923	3.991	0.081	68.819	0.078	
05	U2	NO	36.283	69.365	3.999	0.165	68.899	0.158	$\sigma_e^2 = 4.0$
06	U2	EX	35.765	68.116	4.008	0.572	69.299	0.558	
07	U3	U1	36.106	74.014	4.009	0.082	75.646	0.078	
08	U3	NO	36.114	77.586	4.002	0.168	75.726	0.158	$\rho = 0.8889$
09	U3	EX	36.033	76.571	3.978	0.517	76.126	0.558	
10	U6	U1	35.856	81.605	3.991	0.081	82.473	0.078	
11	U6	NO	36.107	84.019	4.006	0.171	82.553	0.158	$E(MS_A) = 36.0$
12	U6	EX	36.016	82.009	3.992	0.559	82.953	0.558	
13	NO	U1	35.931	81.749	4.005	0.084	89.299	0.078	
14	NO	NO	36.016	85.815	3.998	0.162	89.379	0.158	$E(MS_e) = 4.0$
15	NO	EX	36.130	90.371	3.971	0.523	89.779	0.558	
16	EX	U1	35.335	258.850	4.005	0.085	294.099	0.078	
17	EX	NO	35.358	270.874	3.998	0.162	294.179	0.158	
18	EX	EX	35.380	269.880	3.972	0.523	294.579	0.558	
19	U1	U1	11.924	7.389	3.994	0.084	7.291	0.078	
20	U1	NO	12.028	7.147	3.992	0.168	7.371	0.158	$\sigma_A^2 = 1.0$
21	U1	EX	12.030	7.657	3.981	0.557	7.771	0.558	
22	U2	U1	12.011	8.902	3.999	0.082	8.571	0.078	
23	U2	NO	11.982	8.550	3.991	0.164	8.651	0.158	$\sigma_e^2 = 4.0$
24	U2	EX	12.047	8.953	4.021	0.568	9.051	0.558	
25	U3	U1	11.928	9.132	3.996	0.084	8.998	0.078	
26	U3	NO	11.945	9.455	4.013	0.158	9.078	0.158	$\rho = 0.6667$
27	U3	EX	11.933	9.048	3.993	0.538	9.478	0.558	
28	U6	U1	12.051	9.794	4.000	0.084	9.424	0.078	
29	U6	NO	12.033	9.430	3.993	0.165	9.504	0.158	$E(MS_A) = 12.0$
30	U6	EX	11.954	9.972	3.979	0.536	9.904	0.558	
31	NO	U1	12.050	9.556	3.994	0.084	9.851	0.078	
32	NO	NO	12.029	9.492	3.992	0.168	9.931	0.158	$E(MS_e) = 4.0$
33	NO	EX	12.076	10.484	3.981	0.557	10.331	0.558	
34	EX	U1	12.005	23.122	3.994	0.085	22.651	0.078	
35	EX	NO	11.883	22.484	3.992	0.168	22.731	0.158	
36	EX	EX	11.913	23.485	3.981	0.557	23.131	0.558	
37	U1	U1	6.875	2.910	3.991	0.087	2.853	0.078	
38	U1	NO	6.869	2.886	4.006	0.171	2.933	0.158	$\sigma_A^2 = 0.36$
39	U1	EX	6.868	3.180	3.997	0.526	3.333	0.558	
40	U2	U1	6.908	3.026	3.997	0.084	3.019	0.078	
41	U2	NO	6.805	3.162	3.987	0.160	3.099	0.158	$\sigma_e^2 = 4.0$
42	U2	EX	6.849	3.376	4.002	0.536	3.499	0.558	
43	U3	U1	6.939	3.394	3.980	0.087	3.074	0.078	
44	U3	NO	6.853	3.111	3.989	0.162	3.154	0.158	$\rho = 0.4186$
45	U3	EX	6.805	3.522	3.982	0.550	3.554	0.558	
46	U6	U1	6.943	3.115	4.000	0.082	3.129	0.078	
47	U6	NO	6.907	3.240	3.985	0.158	3.209	0.158	$E(MS_A) = 6.88$
48	U6	EX	6.816	3.476	3.998	0.519	3.609	0.558	
49	NO	U1	6.911	3.161	3.991	0.087	3.184	0.078	
50	NO	NO	6.883	3.362	4.006	0.171	3.264	0.158	$E(MS_e) = 4.0$
51	NO	EX	6.858	3.629	3.997	0.526	3.664	0.558	
52	EX	U1	6.862	4.534	3.991	0.087	4.843	0.078	
53	EX	NO	6.865	4.627	4.006	0.171	4.923	0.158	
54	EX	EX	6.854	4.995	3.997	0.526	5.323	0.558	

(5.6)

(a) $$\mathrm{Var}\,(MS_A) = \left[\frac{2}{I-1} + \frac{1}{I}\left\{\rho^2\gamma_A + (1-\rho)^2\,\gamma_e/J\right\}\right](J\sigma_A^2 + \sigma_e^2)^2$$

(b) $$\mathrm{Var}\,(MS_e) = \left[\frac{2}{(I-1)(J-1)} + \frac{\gamma_e}{IJ}\right]\sigma_e^4$$

TABLE 5.2

Comparisons of Observed Means and Standard Errors of Reliability Estimates Under ANOVA Model and Various Combinations of True and Error Score Distributions With the Values Obtainable From Formulas (5.3), (5.1)-(b), and (5.10), N = 2000, I = 30, J = 8

Ex. No.	Dis Tr.	Er.	Observed β Mean (1)	S.E. (2)	Calculated S.E. (β) by (5.3) (3)	(5.1)-(b) (4)	(5.10) (5)	Parameters and $E(\beta)$ by (5.1)-(a) (6)
01	U1	U1	0.884	0.028	0.038	0.036	0.030	
02	U1	NO	0.885	0.028	0.038	0.036	0.030	$\sigma_A^2 = 4.0$
03	U1	EX	0.886	0.033	0.038	0.036	0.030	
04	U2	U1	0.884	0.031	0.038	0.036	0.033	
05	U2	NO	0.833	0.032	0.038	0.036	0.033	$\sigma_e^2 = 4.0$
06	U2	EX	0.882	0.038	0.038	0.036	0.033	
07	U3	U1	0.882	0.033	0.038	0.036	0.034	
08	U3	NO	0.882	0.038	0.038	0.036	0.034	$\rho = 0.8889$
09	U3	EX	0.883	0.038	0.038	0.036	0.034	
10	U6	U1	0.881	0.036	0.038	0.036	0.035	
11	U6	NO	0.881	0.035	0.038	0.036	0.035	$E(\beta) = 0.8807$
12	U6	EX	0.882	0.040	0.038	0.026	0.035	
13	NO	U1	0.881	0.035	0.038	0.036	0.036	
14	NO	NO	0.881	0.036	0.038	0.036	0.036	
15	NO	EX	0.882	0.040	0.038	0.036	0.036	
16	EX	U1	0.864	0.062	0.038	0.036	0.057	
17	EX	NO	0.864	0.061	0.038	0.036	0.057	
18	EX	EX	0.865	0.064	0.038	0.036	0.057	
19	U1	U1	0.646	0.090	0.101	0.108	0.098	
20	U1	NO	0.650	0.091	0.101	0.108	0.098	$\sigma_A^2 = 1.0$
21	U1	EX	0.653	0.098	0.101	0.108	0.100	
22	U2	U1	0.644	0.103	0.101	0.108	0.103	
23	U2	NO	0.646	0.100	0.101	0.108	0.103	$\sigma_e^2 = 4.0$
24	U2	EX	0.647	0.107	0.101	0.108	0.105	
25	U3	U1	0.642	0.101	0.101	0.108	0.104	
26	U3	NO	0.640	0.108	0.101	0.108	0.105	$\rho = 0.6667$
27	U3	EX	0.645	0.107	0.101	0.108	0.106	
28	U6	U1	0.643	0.106	0.101	0.108	0.106	
29	U6	NO	0.645	0.103	0.101	0.108	0.106	$E(\beta) = 0.6420$
30	U6	EX	0.645	0.111	0.101	0.108	0.108	
31	NO	U1	0.644	0.106	0.101	0.108	0.108	
32	NO	NO	0.645	0.105	0.101	0.108	0.108	
33	NO	EX	0.647	0.114	0.101	0.108	0.109	
34	EX	U1	0.615	0.154	0.101	0.108	0.146	
35	EX	NO	0.612	0.156	0.101	0.108	0.147	
36	EX	EX	0.617	0.158	0.101	0.108	0.148	
37	U1	U1	0.380	0.175	0.151	0.188	0.180	
38	U1	NO	0.380	0.172	0.151	0.188	0.182	$\sigma_A^2 = 0.36$
39	U1	EX	0.384	0.179	0.151	0.188	0.189	
40	U2	U1	0.380	0.181	0.151	0.188	0.183	
41	U2	NO	0.369	0.195	0.151	0.188	0.185	$\sigma_e^2 = 4.0$
42	U2	EX	0.379	0.185	0.151	0.188	0.192	
43	U3	U1	0.381	0.189	0.151	0.188	0.184	
44	U3	NO	0.376	0.186	0.151	0.188	0.186	$\rho = 0.4186$
45	U3	EX	0.376	0.189	0.151	0.188	0.193	
46	U6	U1	0.382	0.181	0.151	0.188	0.185	
47	U6	NO	0.381	0.184	0.151	0.188	0.187	$E(\beta) = 0.3755$
48	U6	EX	0.375	0.187	0.151	0.188	0.194	
49	NO	U1	0.379	0.185	0.151	0.188	0.186	
50	NO	NO	0.374	0.188	0.151	0.188	0.188	
51	NO	EX	0.379	0.185	0.151	0.188	0.195	
52	EX	U1	0.360	0.213	0.151	0.188	0.216	
53	EX	NO	0.357	0.219	0.151	0.188	0.217	
54	EX	EX	0.362	0.217	0.151	0.188	0.224	

$$(5.1)\quad \begin{cases} \text{(a)} & E(\beta) = 1 - (1-\rho)\dfrac{I-1}{I-3} \\ \text{(b)} & \text{Var}(\beta) = (1-\rho)^2\,\dfrac{2(I-1)(\nu+I-3)}{(J-1)(I-3)^2(I-5)} \end{cases}$$

$$(5.3)\quad \text{Var}(\beta) = \frac{(1-\rho^2)^2}{I}$$

$$(5.10)\quad \text{Var}(\beta) \simeq (1-\rho)^2\left[\frac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \frac{\gamma_y}{I}\right]$$

almost negligible, as expected from the discussion of the standard errors in Section 5.1.3. Therefore, as far as point estimation is concerned, negative kurtosis would not cause any serious problems, but positive kurtosis may cause serious underestimation of reliability.

Table 5.2 also contains the standard errors of $\hat{\rho}$ in columns (3), (4), and (5) under various formulas as well as the observed results in column (2). The results clearly indicate the inappropriateness of the tranditional formula (5.3) or the more recent formula (5.1)-(b) when γ_A is non-zero, and demonstrates the effectiveness of formula (5.10). To see how closely the values based on these formulas approximate the observed values, the sum of squares of the deviation from the observed values are calculated with the results 0.0416, 0.0133, and 0.0013 for formulas (5.3), (5.1)-(b) and (5.10) respectively, with the minimum deviations for formula (5.10).

To examine the robustness of the F-test based on formula (2.17), under normal distribution theory, the shapes of the upper and lower 5% tail portions of the distributions of $\hat{\rho}$ were investigated. Columns (1) and (2) of Table 5.3 show approximate real Type one errors when nominal significance levels are fixed at 5% level for each tail. The results clearly indicate that real Type one errors are less than the nominal value if γ_A is negative, and the smaller is γ_A, the smaller is the resulting real Type one error. For positive γ_A, the real Type one errors are greater than the nominal value. These results are in close agreement with the Scheffé's conclusion referred to in Section 5.1.2. It is also noticed that the effect of non-zero γ_A is less for small ρ, i.e., the test is robust if ρ tends to zero as

TABLE 5.3

Comparisons of Observed Lower and Upper 5% Critical Points of Reliability Estimates Under the ANOVA Model Using Various Combinations of True and Error Score Distributions, and Real Type One Errors of F-Test When Nominal Value is 5% With the Values Obtainable Under the Normal Theory, N = 2000, I = 30, J = 8

Ex. No.	Dis. Tr.	Er.	Real Sig. (%) Lower (1)	Upper (2)	Observed C.P.[1] Lower (3)	Upper (4)	Theoretical C.P.[2] Lower (5)	Upper (6)	Parameters[3] (7)
01	U1	U1	1.80	1.65	0.837	0.919	0.814	0.927	
02	U1	NO	1.80	1.80	0.838	0.920	0.814	0.927	$\sigma_A^2 = 4.0$
03	U1	EX	2.25	5.05	0.829	0.927	0.814	0.927	
04	U2	U1	2.75	2.45	0.827	0.922	0.814	0.927	
05	U2	NO	3.05	3.45	0.827	0.924	0.814	0.927	$\sigma_e^2 = 4.0$
06	U2	EX	4.85	6.00	0.815	0.929	0.814	0.927	
07	U3	U1	3.30	3.05	0.823	0.923	0.814	0.927	
08	U3	NO	4.15	4.40	0.821	0.926	0.814	0.927	$\rho = 0.8889$
09	U3	EX	5.00	7.25	0.813	0.930	0.814	0.927	
10	U6	U1	4.30	3.60	0.816	0.924	0.814	0.927	
11	U6	NO	4.65	4.70	0.816	0.926	0.814	0.927	
12	U6	EX	5.60	7.20	0.812	0.930	0.814	0.927	
13	NO	U1	4.70	3.90	0.816	0.925	0.814	0.927	
14	NO	NO	4.25	4.85	0.819	0.926	0.814	0.927	
15	NO	EX	5.30	8.30	0.811	0.932	0.814	0.927	
16	EX	U1	17.40	10.50	0.747	0.940	0.814	0.927	
17	EX	NO	17.35	11.15	0.744	0.942	0.814	0.927	
18	EX	EX	17.95	12.90	0.752	0.944	0.814	0.927	
19	U1	U1	3.35	3.00	0.474	0.769	0.442	0.781	
20	U1	NO	3.00	3.45	0.483	0.772	0.442	0.781	$\sigma_A^2 = 1.0$
21	U1	EX	3.50	5.40	0.471	0.783	0.442	0.781	
22	U2	U1	4.55	3.85	0.450	0.775	0.442	0.781	
23	U2	NO	3.95	4.00	0.470	0.776	0.442	0.781	$\sigma_e^2 = 4.0$
24	U2	EX	4.65	6.20	0.448	0.788	0.442	0.781	
25	U3	U1	4.15	3.85	0.453	0.774	0.442	0.781	
26	U3	NO	5.05	4.45	0.441	0.778	0.442	0.781	$\rho = 0.6667$
27	U3	EX	4.80	5.45	0.448	0.785	0.442	0.781	
28	U6	U1	4.45	4.45	0.446	0.777	0.442	0.781	
29	U6	NO	4.75	4.30	0.445	0.778	0.442	0.781	
30	U6	EX	5.35	6.65	0.440	0.788	0.442	0.781	
31	NO	U1	4.55	4.80	0.448	0.780	0.442	0.781	
32	NO	NO	4.55	5.05	0.449	0.782	0.442	0.781	
33	NO	EX	5.95	7.25	0.427	0.792	0.442	0.781	
34	EX	U1	11.55	10.55	0.327	0.814	0.442	0.781	
35	EX	NO	11.95	11.45	0.331	0.813	0.442	0.781	
36	EX	EX	12.80	12.10	0.321	0.822	0.442	0.781	
37	U1	U1	3.80	5.05	0.056	0.619	0.027	0.618	
38	U1	NO	3.75	4.50	0.070	0.612	0.027	0.618	$\sigma_A^2 = 0.36$
39	U1	EX	4.65	4.75	0.040	0.616	0.027	0.618	
40	U2	U1	4.65	5.25	0.034	0.622	0.027	0.618	
41	U2	NO	5.50	4.65	0.070	0.615	0.027	0.618	$\sigma_e^2 = 4.0$
42	U2	EX	5.10	5.35	0.025	0.623	0.027	0.618	
43	U3	U1	5.05	5.65	0.026	0.626	0.027	0.618	
44	U3	NO	5.05	4.55	0.026	0.615	0.027	0.618	$\rho = 0.4186$
45	U3	EX	4.85	5.55	0.033	0.618	0.027	0.618	
46	U6	U1	4.30	4.85	0.053	0.615	0.027	0.618	
47	U6	NO	4.55	4.85	0.041	0.617	0.027	0.618	
48	U6	EX	4.55	4.65	0.043	0.614	0.027	0.618	
49	NO	U1	4.60	4.60	0.033	0.616	0.027	0.618	
50	NO	NO	5.10	4.55	0.023	0.615	0.027	0.618	
51	NO	EX	4.70	4.85	0.029	0.618	0.027	0.618	
52	EX	U1	7.30	7.05	-0.026	0.637	0.027	0.618	
53	EX	NO	8.25	7.10	-0.049	0.644	0.027	0.618	
54	EX	EX	7.35	8.65	-0.049	0.645	0.027	0.618	

[1] Observed lower and upper 5% critical points of $\hat{\rho}$.

[2] Theoretical lower and upper 5% critical points of $\hat{\rho}$ with normal distribution of true and error scores.

[3] Kurtosis of the random variables U1, U2, U3, U6, NO, and EX are given in Table 4.3.

anticipated by the earlier discussion of the standard errors of estimation in Section 5.1.3.

5.1.5 Conclusions on the Effects of Non-Normality Under ANOVA Model

From the above discussions the following conclusions are tentatively made:

(a) The effect of non-normality of the error score distribution is negligible for J fairly large, where J is the number of part-tests.

(b) Non-zero kurtosis of the true score distribution substantially effects the sampling distribution and standard error of reliability estimates.

(c) The F-test under normal theory is robust for near zero population reliability, or near zero true score kurtosis, if J is fairly large.

(d) Formula (5.10) is superior to the traditional formula (5.3) or (5.1)-(b) for the calculation of the standard error of reliability estimates.

(e) For the F-test, the real Type one error is lower than the nominal value for negative kurtosis, and higher for positive kurtosis of true scores. This true score kurtosis is closely approximated by test score kurtosis divided by the square of the reliability.

The above findings are restricted to the ANOVA model, and generalization to more liberal test score models requires further study.

5.2.0 Relaxation of the Homogeneity of Error Variance Constraint in the ANOVA Model

5.2.1 The ETEM Model

For the ANOVA model it was assumed that the variances of error scores $\{e_{ij}\}$ were homogeneous, i.e., all the error variances $\{\sigma^2_{ej}\}$ are equal to an unknown constant σ^2_e by assumption (2.7)-(d). This assumption was made not because real data are expected to have homogeneous error variances, but to make the mathematical abstraction simpler. Therefore it is conceivable that the error variances may differ for each part test, i.e., for real data the variance of e_{ij} may depend on the part test j, as given by (2.19). Under this last assumption, the model becomes an essentially τ equivalent measurement (ETEM) which was discussed more fully in Chapter Two. Under this model, there is not a common intra-class correlation among the J part-scores to be interpreted as the reliability of a part-test under the ANOVA model. But the reliability is still equal to the Alpha coefficient. The only difference from the ANOVA model is the replacement of σ^2_e in the reliability formula (2.12) by the mean of $\{\sigma^2_{ej}\}$, denoted by $\sigma^2_{e.}$.

Because assumption (2.7)-(d) is violated, the distribution of reliability estimates given by (2.17) cannot be expected to hold for the ETEM models; at best it is hoped that the distribution is closely approximated or the distribution is robust against the violation of the assumption of homogeneity of error variances.

5.2.2 Effects of Non-Homogeneous Error Variances Assuming Normal Distribution

The general distributional theory of reliability estimates under the ETEM model with the normal assumption is not yet known except

for the case of $J = 2$. Kristof (1970) has shown that, for $J = 2$, the statistic

$$(5.12) \qquad t = \frac{\rho - \hat{\rho}}{\hat{\rho}(1-\rho)^{\frac{1}{2}}} \frac{s_{12}}{(s_1^2\, s_2^2 - s_{12}^2)^{\frac{1}{2}}} (I-2)^{\frac{1}{2}}$$

is distributed as Student's t-statistic with $I-2$ degrees of freedom, where s_1^2, s_2^2, and s_{12} are the sample variances of two part-tests and the covariance between them respectively. Kristof derived this formula by the maximum likelihood method under bivariate normal assumptions for the alpha coefficient, but the formula can also be used interchangeably for the reliability coefficient under the ETEM model.

For the general case, $J > 2$, nothing is known yet, and at present the simulation method provides the only way to investigate the sampling distribution of reliability estimates. Because equation (2.17) does not involve the error variance parameter directly, it may be hoped that the distribution given by (2.17) is still valid or approximately true under the ETEM model if the normality assumptions are not violated. In other words, it is hoped that the distribution is robust against the violation of the homogeneity of error variances assumption to enable the test theorists to use the results obtained under the ANOVA model.

To separate the effect of non-normality from that of non-homogeneous error variances, the ETEM model is first investigated using normal distributions of true and error scores. In order to make comparisons possible, the constants and parameters used for the cases of the ANOVA model are retained except for the values of the error

variances. With 3 levels of σ_A^2, as under the ANOVA model, 6 different sets of non-homogeneous error variances are used for the simulation experiments. The sets of error variances are given in the following table including the homogeneous case (EV1) used under the ANOVA model as a special case.

TABLE 5.4

Summary of Error Variances used Under ETEM Model

Notation	Error Variances $\{\sigma_{ej}^2\}$								Mean $\sigma_{e.}^2$	Variance $(\sum(\sigma_{ej}^2 - \sigma_{e.}^2)^2)/J$
	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8		
EV1	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.0000	0.0000
EV2	1.46	2.56	3.24	4.00	4.00	4.84	5.76	6.15	4.0018	2.1887
EV3	1.00	1.00	4.00	9.00	9.00	4.00	1.00	1.00	3.7500	10.6875
EV4	9.00	4.00	1.00	0.25	0.25	1.00	4.00	9.00	3.5625	11.8242
EV5	16.00	9.00	4.00	1.00	1.00	1.00	1.00	1.00	4.2500	26.6875
EV6	1.00	4.00	16.00	16.00	9.00	4.00	1.00	0.00	6.3750	37.7344
EV7	1.00	1.00	16.00	16.00	1.00	1.00	1.00	1.00	4.7500	42.1875

The last two columns of Table 5.4 give $\sigma_{e.}^2$, which is equal to $E(MS_e)$, and the variance of $\{\sigma_{ej}^2\}$ within each set over $J = 8$. To make comparisons easy, these sets are ordered with increasing degree of non-homogeneity, measured by the variance within each set, which has a range of 0.0 to 42.1875.

Table 5.5 summarizes the mean and variance MS_A and MS_e, for $N = 2000$ samples in columns (1) to (4) inclusive, and compares the results with those obtainable from formula (5.6) with σ_e^2

TABLE 5.5

Comparisons of Observed Means and Variances of MS's Under the ETEM Model and Normal Distributions With the Values Obtainable From Formula (5.6), N = 2000, I = 30, J = 8

Ex. No.	Error Set	Observed MS_A Mean (1)	Observed MS_A Var. (2)	Observed MS_e Mean (3)	Observed MS_e Var. (4)	$E(MS_A)$ (5)	Var. by (5.6) MS_A (6)	Var. by (5.6) MS_e (7)	Parameters (8)
01	EV1	36.016	85.815	3.998	0.162	36.000	89.379	0.158	
02	EV2	36.041	85.556	4.000	0.184	36.002	89.388	0.158	$\sigma_A^2 = 4.0$
03	EV3	35.727	88.222	3.742	0.226	35.750	88.142	0.139	
04	EV4	35.879	98.868	3.543	0.242	35.562	87.220	0.125	
05	EV5	36.266	86.094	4.244	0.387	36.250	90.625	0.178	
06	EV6	38.049	101.514	6.378	0.742	38.375	101.561	0.400	
07	EV7	36.784	94.840	4.705	0.551	36.750	93.142	0.222	
08	EV1	12.029	9.492	3.992	0.168	12.000	9.931	0.158	
09	EV2	12.034	9.581	3.995	0.189	12.002	9.934	0.158	$\sigma_A^2 = 1.0$
10	EV3	11.898	9.537	3.740	0.236	11.750	9.522	0.139	
11	EV4	11.495	9.317	3.549	0.220	11.562	9.220	0.125	
12	EV5	12.278	10.465	4.236	0.407	12.250	10.349	0.178	
13	EV6	14.563	14.483	6.366	0.753	14.375	14.251	0.400	
14	EV7	12.838	11.499	4.780	0.603	12.750	11.211	0.222	
15	EV1	6.883	3.362	4.006	0.171	6.880	3.264	0.158	
16	EV2	6.891	3.411	4.008	0.189	6.882	3.266	0.158	$\sigma_A^2 = 0.36$
17	EV3	6.700	3.203	3.762	0.221	6.630	3.032	0.139	
18	EV4	6.430	2.779	3.559	0.230	6.442	2.862	0.125	
19	EV5	7.120	3.491	4.252	0.429	7.130	3.506	0.178	
20	EV6	9.341	6.174	6.403	0.704	9.255	5.907	0.400	
21	EV7	7.629	4.035	4.750	0.619	7.630	4.015	0.222	

$E(MS_e) = \sigma_{e.}^2 =$ 4.000 (EV1)
4.002 (EV2)
3.750 (EV3)
3.563 (EV4)
4.250 (EV5)
6.375 (EV6)
4.750 (EV7)

$$(5.6)\quad \begin{cases} \text{(a)} \quad \mathrm{Var}\,(MS_A) = [\frac{2}{I-1} + \frac{1}{I}\{\rho^2\gamma_A + (1-\rho)^2\,\gamma_e/J\}](J\sigma_A^2 + \sigma_e^2)^2 \\ \text{(b)} \quad \mathrm{Var}\,(MS_e) = [\frac{2}{(I-1)(J-1)} + \frac{\gamma_e}{IJ}]\,\sigma_e^4 \end{cases}$$

replaced by $\sigma^2_{e.}$, and $\gamma_A = \gamma_e = 0$ in columns (6) and (7). Close agreement between expected MS's and the mean of observed MS's is seen as was the case for the ANOVA model. More specifically, the expected and observed variance of MS_A, columns (2) and (6), agree closely, but the observed variance of MS_e, column (4), differs greatly from the theoretical value obtainable from (5.6) given in column (7). The greater the non-homogeneity of error variances, the greater the discrepancy noted, reaching in the extreme a factor of three for experiment 21. Therefore, it may be concluded that the formula (5.6) cannot be applied blindly in the case of the ETEM model, due to the possible effect of non-homogeneity of error variances.

Table 5.6 summarizes the observed mean and standard error for each experiment in columns (3) and (4) and compares it with the values obtainable from (5.1), (5.3), and (5.10) given in columns (2), (5), and (6). It is observed that a rather close agreement exists between the observed mean of $\hat{\rho}$ and $E(\hat{\rho})$ obtainable from (5.1)-(a) under the ANOVA model, i.e., columns (2) and (3), indicating robustness of the ETEM model as far as point estimation and biasedness are concerned. For the standard error of estimation, all two formulas predict the observed values reasonably well. Formula (5.10) seems better than (5.3), though the difference is not great. The calculated sum of squares of the deviation from the observed values are 0.00858 and 0.00097 for formulas (5.3) and (5.10) respectively, confirming the conclusion. All of these results suggest that the standard error of reliability estimate is robust against the violation of homogeneity of error variances.

TABLE 5.6

Comparisons of Observed Means and Standard Errors of Reliability Estimates Under ETEM Model and Normal Distributions With the Values Obtainable From Formula (5.3), and (5.10), N = 2000, I = 30, J = 8

No.	Error Set	Rel. (1)	$E(\hat{\rho})$ by (5.1)-(a) (2)	Observed $\hat{\rho}$ Mean (3)	S.E. (4)	S.E. by formulas (5.3) (5)	(5.10) (6)	Parameters (7)
01	EV1	0.889	0.881	0.881	0.036	0.038	0.036	
02	EV2	0.889	0.881	0.881	0.036	0.038	0.036	$\sigma_A^2 = 4.0$
03	EV3	0.895	0.887	0.887	0.036	0.036	0.034	
04	EV4	0.899	0.892	0.893	0.035	0.035	0.032	
05	EV5	0.883	0.874	0.875	0.039	0.040	0.038	
06	EV6	0.834	0.822	0.820	0.058	0.056	0.054	
07	EV7	0.871	0.861	0.863	0.044	0.044	0.042	
08	EV1	0.667	0.642	0.645	0.105	0.101	0.108	
09	EV2	0.667	0.642	0.644	0.106	0.102	0.108	$\sigma_A^2 = 1.0$
10	EV3	0.681	0.657	0.664	0.099	0.098	0.103	
11	EV4	0.692	0.669	0.669	0.102	0.095	0.110	
12	EV5	0.653	0.627	0.631	0.114	0.105	0.112	
13	EV6	0.557	0.524	0.533	0.127	0.126	0.143	
14	EV7	0.628	0.600	0.603	0.119	0.111	0.121	
15	EV1	0.419	0.376	0.374	0.188	0.151	0.188	
16	EV2	0.419	0.375	0.375	0.187	0.151	0.188	$\sigma_A^2 = 0.36$
17	EV3	0.434	0.393	0.398	0.180	0.148	0.183	
18	EV4	0.447	0.406	0.410	0.173	0.146	0.179	
19	EV5	0.404	0.360	0.365	0.184	0.153	0.193	
20	EV6	0.311	0.260	0.267	0.201	0.165	0.223	
21	EV7	0.378	0.331	0.338	0.197	0.157	0.201	

(5.1)-(a) $\quad E(\hat{\rho}) = 1 - (1-\rho)\dfrac{I-1}{I-3}$

(5.3) $\quad \mathrm{Var}\ (\hat{\rho}) = \dfrac{(1-\rho^2)^2}{I}$

(5.10) $\quad \mathrm{Var}\ (\hat{\rho}) \simeq (1-\rho)^2\ \left[\dfrac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \dfrac{\gamma_y}{I}\right]$.

Table 5.7 summarizes the lower and upper 5% portions of the observed distribution of $\hat{\rho}$ in columns (2) and (3) and compares them with the values obtainable under the ANOVA model and normal theory given in columns (4) and (5), namely from formula (2.17). The table also gives approximate real Type one error in columns (6) and (7) when the F-test of (2.17) is used for the ETEM model with normal distributions. The results clearly indicate the robustness of the F-test against the violation of homogeneity of error variance assumptions. Although there is a case (experiment 4) which gives as much as an 8% level of Type one error, there seems to be no systematic inflation or deflation of the nominal Type one error as a whole.

5.2.3 Effects of Non-Normality on ETEM Model

In the previous section, it was seen that the effect of non-homogeneous error variances on sampling distribution of reliability estimates is minimal, and it was also seen in Section 5.1 that the sampling distribution is sensitive only to the violation of the assumption of the normality of true scores and is robust against distributional assumption of error scores. Therefore, it is logical to expect that the distribution is not robust against the distributional assumption of true scores, but the effect of non-normality of error scores must still be investigated under the ETEM model, since there is a possibility of interaction between the non-normal error score distribution and non-homogeneous error variances.

To investigate this interaction effect, further experiments were carried out using the EV2 error variances set, chosen because its $\sigma^2_{e.} = 4.0018$ is closest to $\sigma^2_e = 4.0$ used for the ANOVA model

TABLE 5.7

Comparisons of Observed Lower and Upper 5% Critical Points and Real Type One Errors of F-Test When Nominal Value is Fixed at 5%, Under ETEM Model and Normal Distributions With the Values Obtainable Under ANOVA Model, N = 2000, I = 30, J = 8

No.	Error Set	Rel.	Observed C.P.[1]		Theoretical C.P.[2]		Real Sig. (%)		Parameters
			Lower	Upper	Lower	Upper	Lower	Upper	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
01	EV1	0.889	0.819	0.926	0.814	0.927	4.25	4.85	
02	EV2	0.889	0.818	0.927	0.814	0.927	4.10	4.90	$\sigma_A^2 = 4.0$
03	EV3	0.895	0.821	0.932	0.824	0.931	5.55	6.05	
04	EV4	0.900	0.831	0.940	0.832	0.934	5.25	8.05	
05	EV5	0.883	0.802	0.925	0.804	0.923	5.20	6.00	
06	EV6	0.834	0.719	0.892	0.722	0.891	5.30	5.65	
07	EV7	0.871	0.787	0.920	0.784	0.915	4.65	7.15	
08	EV1	0.667	0.449	0.782	0.442	0.781	4.55	5.05	
09	EV2	0.667	0.442	0.784	0.442	0.782	5.00	5.20	$\sigma_A^2 = 1.0$
10	EV3	0.681	0.479	0.793	0.466	0.791	3.85	5.45	
11	EV4	0.692	0.475	0.801	0.484	0.798	5.45	6.10	
12	EV5	0.653	0.414	0.778	0.419	0.772	5.20	6.25	
13	EV6	0.557	0.292	0.712	0.258	0.709	3.80	5.25	
14	EV7	0.628	0.385	0.756	0.376	0.755	4.50	5.40	
15	EV1	0.419	0.023	0.616	0.027	0.618	5.10	4.55	
16	EV2	0.419	0.018	0.617	0.026	0.618	5.15	4.70	$\sigma_A^2 = 0.36$
17	EV3	0.434	0.057	0.633	0.053	0.629	4.80	5.45	
18	EV4	0.447	0.095	0.627	0.074	0.637	4.55	4.05	
19	EV5	0.404	0.020	0.604	0.002	0.609	4.25	4.45	
20	EV6	0.311	-0.135	0.549	-0.153	0.543	4.35	5.05	
21	EV7	0.378	-0.031	0.582	-0.042	0.591	4.60	4.10	

[1]Observed lower and upper 5% critical points of $\hat{\rho}$.

[2]Theoretical lower and upper 5% critical points of $\hat{\rho}$ under ANOVA model.

to make comparisons simpler, and three levels of σ_A^2, for three types of true and error score distributions, namely uniform (U1), normal (NO), and exponential (EX). Altogether the results of 27 experiments are summarized by tabulating the MS's (Table 5.8), standard errors (Table 5.9), and lower and upper 5% critical points of the distribution of reliability estimates with approximate real Type one errors when the nominal values are fixed at 5% level (Table 5.10).

These 27 experiments may be compared with the results of the corresponding experiments under the ANOVA model, namely experiments 1-3, 13-21, 31-39, and 49-54 of Tables 5.1, 5.2, and 5.3. The expected values of MS's and variance of MS_A show close agreement with observed values, but formula (5.6) does consistently underestimate the variance of MS_e, though the difference is trivial. Table 5.9 suggests that formula (5.10) closely approximates the observed standard error as in the case of ANOVA model. Observation of Table 5.10 also suggests that the pattern of discrepancy of real Type one error from the nominal value of 5% is almost the same as for the case of the ANOVA model, thus indicating non-existence of interaction effects between the non-homogeneous variance and non-normality of error score distributions.

5.2.4 Conclusions for the Distributions Under ETEM Model

The effects of non-homogeneous error variances on the sampling distribution of reliability estimates was investigated by simulating 21 experiments using three levels of ρ and 7 sets of error variances whose variance ranged from 0.0 to 42.1876. The following conclusions are tentatively made.

TABLE 5.8

Comparisons of Observed Means and Variances of MS's Under ETEM Model with EV2 Error Variances Set and Various Combinations of True and Error Score Distributions with the Values Obtainable from Formula (5.6), $N = 2000,\ I = 30,\ J = 8$

Dis			Observed MS_A		Observed MS_e		Var. by (5.6)		Parameters, Expected Values Under ANOVA
No.	Tr.	Er.	Mean	Var.	Mean	Var.	MS_A	MS_e	
			(1)	(2)	(3)	(4)	(5)	(6)	(7)
01	UI	UI	36.063	46.826	4.005	0.094	48.348	0.078	$\sigma_A^2 = 4.0$
02	UI	NO	36.119	45.750	3.996	0.184	48.428	0.158	
03	UI	EX	36.041	45.461	3.971	0.597	48.829	0.558	$\rho = 0.8888$
04	NO	UI	35.946	81.623	4.005	0.094	89.308	0.078	$E(MS_A) = 36.0$
05	NO	NO	36.041	85.556	4.000	0.184	89.388	0.158	
06	NO	EX	36.112	90.902	3.971	0.597	89.789	0.558	$E(MS_e) = 4.0018$
07	EX	UI	35.328	259.119	4.005	0.094	294.108	0.078	
08	EX	NO	35.347	269.983	3.996	0.184	294.188	0.158	
09	EX	EX	35.378	269.587	3.971	0.597	294.589	0.558	
10	UI	UI	11.913	7.355	3.999	0.092	7.294	0.078	$\sigma_A^2 = 1.0$
11	UI	NO	12.019	7.240	3.995	0.189	7.374	0.158	
12	UI	EX	12.036	7.769	3.985	0.632	7.774	0.558	$\rho = 0.6666$
13	NO	UI	12.049	9.579	3.999	0.092	9.854	0.078	$E(MS_A) = 12.0018$
14	NO	NO	12.034	9.581	3.995	0.189	9.934	0.158	
15	NO	EX	12.083	10.585	3.985	0.632	10.334	0.558	
16	EX	UI	12.012	22.965	3.999	0.093	22.654	0.078	$E(MS_e) = 4.0018$
17	EX	NO	11.901	22.794	3.995	0.189	22.734	0.158	
18	EX	EX	11.918	23.848	3.985	0.632	23.134	0.558	
19	UI	UI	6.877	2.916	3.990	0.097	2.854	0.078	$\sigma_A^2 = 0.36$
20	UI	NO	6.879	2.891	4.008	0.189	2.934	0.158	
21	UI	EX	6.886	3.264	4.002	0.607	3.335	0.558	$\rho = 0.4185$
22	NO	UI	6.912	3.184	3.990	0.097	3.186	0.078	
23	NO	NO	6.891	3.411	4.008	0.189	3.266	0.158	$E(MS_A) = 6.8818$
24	NO	EX	6.870	3.718	4.002	0.607	3.667	0.558	
25	EX	UI	6.870	4.620	3.990	0.097	4.845	0.078	$E(MS_e) = 4.0018$
26	EX	NO	6.873	4.662	4.008	0.189	4.925	0.158	
27	EX	EX	6.858	5.080	4.002	0.607	5.325	0.558	

$$
(5.6)\quad \begin{cases} \text{(a)}\quad Var\ (MS_A) = [\frac{2}{I-1} + \frac{1}{I}\{\rho^2\gamma_A + (1-\rho)^2\gamma_e/J\}](J\ \sigma_A^2 + \sigma_e^2)^2 \\[2ex] \text{(b)}\quad Var\ (MS_e) = [\frac{2}{(I-1)(J-1)} + \frac{\gamma_e}{IJ}]\ \sigma_e^4 \end{cases}
$$

TABLE 5.9

Comparisons of Observed Means and Standard Errors of Reliability Estimates Under ETEM Model With EV2 Error Variances Set and Various Combinations of True and Error Score Distributions With the Values Obtainable From Formulas (5.3), (5.1), and (5.10), N = 2000, I = 30, J = 8

No.	Dis. Tr.	Er.	Observed $\hat{\rho}$ Mean (1)	S.E. (2)	Calculated from formulas (5.3) (3)	(5.1) (4)	(5.10) (5)	Parameters (6)
01	U1	U1	0.884	0.028	0.038	0.036	0.030	
02	U1	NO	0.885	0.028	0.038	0.036	0.030	$\sigma_A^2 = 4.0$
03	U1	EX	0.886	0.033	0.038	0.036	0.030	
04	NO	U1	0.881	0.036	0.038	0.036	0.036	$\rho = 0.8888$
05	NO	NO	0.881	0.036	0.038	0.036	0.036	$E(\hat{\rho}) = 0.8806$
06	NO	EX	0.882	0.041	0.038	0.036	0.036	
07	EX	U1	0.863	0.063	0.038	0.036	0.057	
08	EX	NO	0.864	0.062	0.038	0.036	0.057	
09	EX	EX	0.865	0.064	0.038	0.036	0.057	
10	U1	U1	0.646	0.091	0.101	0.108	0.098	
11	U1	NO	0.650	0.092	0.101	0.108	0.098	$\sigma_A^2 = 1.0$
12	U1	EX	0.653	0.101	0.101	0.108	0.100	
13	NO	U1	0.644	0.105	0.101	0.108	0.108	$\rho = 0.6666$
14	NO	NO	0.648	0.114	0.101	0.108	0.109	$E(\hat{\rho}) = 0.6419$
15	NO	EX	0.648	0.114	0.101	0.108	0.109	
16	EX	U1	0.615	0.154	0.101	0.108	0.146	
17	EX	NO	0.639	0.144	0.101	0.108	0.147	
18	EX	EX	0.617	0.158	0.101	0.108	0.148	
19	U1	U1	0.380	0.175	0.151	0.188	0.180	
20	U1	NO	0.381	0.170	0.151	0.188	0.182	$\sigma_A^2 = 0.36$
21	U1	EX	0.386	0.179	0.151	0.188	0.189	
22	NO	U1	0.380	0.184	0.151	0.188	0.187	$\rho = 0.4185$
23	NO	NO	0.375	0.187	0.151	0.188	0.188	$E(\hat{\rho}) = 0.3754$
24	NO	EX	0.380	0.187	0.151	0.188	0.195	
25	EX	U1	0.360	0.212	0.151	0.188	0.216	
26	EX	NO	0.358	0.218	0.151	0.188	0.217	
27	EX	EX	0.362	0.218	0.151	0.188	0.224	

(5.1) $$\begin{cases} \text{(a)} \quad E(\hat{\rho}) = 1 - (1-\rho)\dfrac{I-1}{I-3} \\ \text{(b)} \quad \mathrm{Var}\ (\hat{\rho}) = (1-\rho)^2 \dfrac{2(I-1)(\nu+I-3)}{(J-1)(I-3)^2(I-5)} \end{cases}$$

(5.3) $$\mathrm{Var}\ (\hat{\rho}) = \frac{(1-\rho^2)^2}{I}$$

(5.10) $$\mathrm{Var}\ (\hat{\rho}) \simeq (1-\rho)^2 \left[\frac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \frac{\gamma_y}{I}\right]$$

TABLE 5.10

Comparisons of Observed Lower and Upper Critical Points of Reliability Estimates and Real Type One Errors of F-Test When Nominal Value is 5%, Under ETEM Model With EV2 Error Variances Set and Various Combinations of True and Error Score Distributions With the Values Obtainable Under the ANOVA Model and Normal Theory, N = 2000, I = 30, J = 8

No.	Dis. Tr.	Er.	Real Sig. % Lower (1)	Real Sig. % Upper (2)	Observed C.P.[1] Lower (3)	Observed C.P.[1] Upper (4)	Theoretical C.P.[2] Lower (5)	Theoretical C.P.[2] Upper (6)	Parameters (7)
01	U1	U1	1.50	1.45	0.838	0.920	0.814	0.927	
02	U1	NO	1.80	1.85	0.839	0.921	0.814	0.927	$\sigma_A^2 = 4.0$
03	U1	EX	3.25	5.85	0.826	0.928	0.814	0.927	
04	NO	U1	4.60	3.95	0.817	0.925	0.814	0.927	$\rho = 0.8888$
05	NO	NO	4.10	4.90	0.818	0.927	0.814	0.927	
06	NO	EX	5.40	8.65	0.810	0.934	0.814	0.927	
07	EX	U1	16.75	10.90	0.746	0.940	0.814	0.927	
08	EX	NO	17.15	11.50	0.745	0.927	0.814	0.927	
09	EX	EX	18.40	12.90	0.746	0.945	0.814	0.927	
10	U1	U1	3.20	3.15	0.478	0.768	0.442	0.781	
11	U1	NO	2.80	3.70	0.481	0.774	0.442	0.781	$\sigma_A^2 = 1.0$
12	U1	EX	3.15	5.25	0.470	0.783	0.442	0.781	
13	NO	U1	4.70	5.05	0.445	0.781	0.442	0.781	
14	NO	NO	5.00	5.20	0.442	0.784	0.442	0.781	$\rho = 0.6666$
15	NO	EX	5.90	7.35	0.422	0.794	0.442	0.781	
16	EX	U1	11.65	9.90	0.337	0.813	0.442	0.781	
17	EX	NO	11.45	11.60	0.326	0.814	0.442	0.781	
18	EX	EX	12.85	11.95	0.323	0.823	0.442	0.781	
19	U1	U1	4.20	4.75	0.050	0.616	0.026	0.618	
20	U1	NO	3.25	4.20	0.072	0.610	0.026	0.618	$\sigma_A^2 = 0.36$
21	U1	EX	4.45	5.10	0.049	0.621	0.026	0.618	
22	NO	U1	4.60	5.30	0.039	0.620	0.026	0.618	
23	NO	NO	5.15	4.70	0.018	0.617	0.026	0.618	$\rho = 0.4185$
24	NO	EX	4.80	5.25	0.033	0.620	0.026	0.618	
25	EX	U1	6.85	7.10	-0.090	0.634	0.026	0.618	
26	EX	NO	7.60	7.10	-0.051	0.647	0.026	0.618	
27	EX	EX	7.20	8.50	-0.037	0.618	0.026	0.618	

[1] Observed lower and upper 5% critical points.

[2] Theoretical lower and upper 5% critical points of $\hat{\rho}$ under ANOVA model with normal distribution of true and error scores.

(a) The variance of MS_e is sensitive to the violation of the homogeneity assumptions, and formula (5.6) should not be used to calculate this statistic.

(b) For the point estimation of reliability, the ANOVA model is quite robust against the violation of homogeneity of error variances provided that the distributions are normal.

(c) The standard error of estimation is quite robust against the violation of the homogeneity of error variances. The best formula is still (5.10).

(d) Formula (2.17) can be used freely without inflating or deflating Type one errors too much for the ETEM model provided that normality is not violated.

The effect of non-normal true or error score distributions under the ETEM model was investigated by performing 27 experiments with three levels of ρ, three types of true and error score distributions, and a set of non-homogeneous error variances. The following conclusions are tentatively made.

(e) Formula (5.6) consistently underestimates the variance of MS_e under the ETEM model.

(f) The interaction between the ETEM model and non-normal error score distribution seems negligible.

(g) The conclusions drawn in Section 5.1.0 may be generalized to the ETEM model with little modification.

5.3.0 <u>Relaxation of the Homogeneity of True Variance Constraint in the ANOVA or ETEM Models</u>

5.3.1 Reliability and the Alpha Coefficient

In Chapter Two, the ANOVA and ETEM models were expanded to include more general models such as the congeneric or multi-factor true score models through the use of the vector or matrix parameters $\underline{\lambda}$ and $\underline{\Lambda}$ in equation (2.6') to produce (2.28). Under these more general models, the ETEM assumptions are not satisfied in general, and the Alpha coefficient is lower than the reliability coefficient. Therefore, one might be interested in two related but different distributions, namely, the sampling distributions of the Alpha coefficient estimates and the reliability estimates. However, the Alpha coefficient has attracted test theorist's interest only because it is considered a practical, and easily computable substitute for the reliability coefficient. Thus, the distribution of the Alpha coefficient estimates is meaningful only in lieu of the distribution of reliability estimates. Furthermore, because no direct estimation formula for reliability is available under these more general models, without exception Alpha coefficient estimates have been accepted as reliability estimates regardless of the underlying models or assumptions.

Test theorists know that the population Alpha coefficient is in general lower than the reliability, but this fact has been frequently confused with underestimation due to biasedness of the estimation procedure. Two kinds of underestimation problems that exist in reliability theory must be distinguished: one is due to deviation from the ETEM assumption, which is not a statistical inference problem, and the other is due to the nature of the estimation formula which is biased.

The sampling distribution of the Alpha coefficient under the

non-ETEM model is the most overlooked aspect of reliability theory. No study has yet been reported on this subject to the author's knowledge. Due to the mathematical complexity involved in these models, it seems almost impossible to investigate the problem by analytical means. Therefore, the problem was investigated as assumption violating cases of the ANOVA model using computer simulation techniques. The major purpose is to find the effects of the violation of the ETEM assumptions, or homogeneity of true score variances and unifactorness of the true score dispersion matrix.

Because so many assumptions of ANOVA models are violated under these more general models, an exhaustive investigation of all the combinations of possible violation of assumptions is prohibitively expensive with the computer simulation method. The study in this section is limited to a few combinations. Therefore, the findings in this section have limited value for generalization.

5.3.2 Distributions Under the Congeneric Model

Under the congeneric true score model, each part-test measures the same trait except for the errors of measurement, i.e., the factorial structure of true scores is unifactor. Therefore all part-test scores have linearly related true scores. Test scores under the classically parallel, ANOVA (or essentially parallel), or ETEM models are all special cases of the congeneric model, as discussed more fully in Chapter Two. In these special cases any true score of a part-test must be essentially identical for a given subject, unlike the congeneric model.

Under the congeneric model, the variance, $\sigma^2_{Aj} = \lambda^2_j$, of true score for part j depends on j, and there is not a common

variance parameter σ_A^2 which has played a key role in the ETEM or ANOVA models. To obtain the corresponding parameters for the congeneric model, a new parameter $\sigma_{A.}^2$ is defined denoting the average of the all elements of the dispersion matrix $\underline{\lambda}\ \underline{\lambda}'$, namely,

$$(5.13) \qquad \sigma_{A.}^2 = (\underline{1}'\ \underline{\lambda}\ \underline{\lambda}'\ \underline{1})/J^2 = (\sum_j \sum_{j'} \lambda_j \lambda_{j'})/J^2 \ .$$

As this parameter is an average of true score variance and covariances, the reliability coefficient is,

$$(5.14) \qquad \rho = \frac{\underline{1}'\ \underline{\lambda}\ \underline{\lambda}'\ \underline{1}}{\underline{1}'(\underline{\lambda}\ \underline{\lambda}' + \underline{\Psi}^2)\underline{1}} = \frac{J^2\ \sigma_{A.}^2}{J^2\ \sigma_{A.}^2 + J\ \sigma_{e.}^2}$$

where $\sigma_{e.}^2$ is the average of error variances as defined by (2.22), namely the mean of $\{\sigma_{ej}^2\}$. Since the distribution (2.17) obtained under the ANOVA model and normal distribution theory does not directly involve the parameters σ_A^2 and σ_e^2, but only directly involves the reliability ρ, it is desirable to know whether the distribution of reliability estimates based on formula (2.13)-(b) is robust against the violation of ETEM assumptions, i.e., whether the relation (2.17) still holds approximately for the congeneric cases.

Under the congeneric model, formula (2.13)-(b) gives the estimate of the Alpha coefficient, not the reliability, but it is hoped that, with moderate violation of ETEM assumptions, inferences based on the estimate of Alpha would not invalidate the inferences of reliability too much as in the case of the previous section.

To see the effects of non-homogeneous true score variances, sampling experiments were performed using the following three sets of $\underline{\lambda}$'s representing three levels of reliability, namely,

$$\underline{\lambda}_1 = \begin{bmatrix} 1.6 \\ 1.8 \\ 1.8 \\ 2.0 \\ 2.0 \\ 2.2 \\ 2.2 \\ 2.4 \end{bmatrix}, \qquad \underline{\lambda}_2 = \begin{bmatrix} 0.8 \\ 0.9 \\ 0.9 \\ 1.0 \\ 1.0 \\ 1.1 \\ 1.1 \\ 1.2 \end{bmatrix}, \qquad \underline{\lambda}_3 = \begin{bmatrix} 0.72 \\ 0.66 \\ 0.66 \\ 0.60 \\ 0.60 \\ 0.54 \\ 0.54 \\ 0.48 \end{bmatrix},$$

which gives three levels of $\sigma^2_{A.}$ namely 4.0, 1.0, and 0.36 and three levels of ρ, i.e., 0.8889, 0.6667, and 0.4186 respectively. The $\underline{\lambda}$'s were chosen such that the values of $\sigma^2_{A.}$ equal σ^2_A used for the ANOVA and ETEM model experiments in Sections 5.1.0 and 5.2.0, in order to facilitate the comparisons. The error variances $\{\sigma^2_{ej}\}$ are fixed at 4.0 as the ANOVA model, and the same constants are used for N, I, and J, i.e., 2000, 30, and 8 respectively. Employing three types of true and error score distributions, namely uniform (U1), normal (NO), and exponential (EX), altogether 27 experiments were performed by REL01, and the results are summarized in Tables 5.11, 5.12, and 5.13. As in the previous sections, the distributions of MS's are examined first. From Table 5.11, it is noted that the effects of non-homogeneous true score variances are minimal, i.e., the results are almost identical with those under ANOVA model given in Table 5.1. Table 5.12 summarizes the means and standard errors of reliability estimates under this model and compares them with the values obtainable from formulas (5.3), (5.1)-(b), and (5.10). It is clearly noticed that formula (5.10) is still the best among the three. When the means of $\hat{\rho}$ in Table 5.12 are compared with the corresponding values of Table 5.2, it may be noticed that under the congeneric model the mean of $\hat{\rho}$ is lower than under the ANOVA model, as expected, since the formula used for the estimation, (2.13), is for the estimation of Alpha, and Alpha is lower than

TABLE 5.11

Comparisons of Observed Means and Variances of MS's Under the Congeneric Model and Various Combinations of True and Error Score Distributions With the Values Obtainable From Formula (5.6), N = 2000, I = 30, J = 8

Ex. No.	Dis. Tr.	Er.	Observed MS_A Mean (1)	Var. (2)	Observed MS_e Mean (3)	Var. (4)	Var. by (5.6) MS_A (5)	MS_e (6)	Parameters and E(MS) (7)
01	U1	U1	36.052	47.112	4.075	0.089	48.339	0.078	$\sigma^2_{A.} = 4.0$
02	U1	NO	36.102	45.563	4.068	0.168	48.419	0.158	
03	U1	EX	36.057	45.036	4.036	0.531	48.819	0.558	
04	NO	U1	35.931	81.750	4.075	0.090	89.299	0.078	$\rho = 0.8889$
05	NO	NO	36.016	85.812	4.069	0.167	89.379	0.158	Alpha = 0.8870
06	NO	EX	36.130	90.374	4.040	0.525	89.779	0.558	$E(MS_A) = 36.0$
07	EX	U1	35.335	258.849	4.071	0.091	294.099	0.078	
08	EX	NO	35.358	270.871	4.064	0.168	294.179	0.158	$E(MS_e) = 4.0$
09	EX	EX	35.380	269.881	4.040	0.525	294.579	0.558	
10	U1	U1	11.924	7.389	4.009	0.085	7.291	0.078	$\sigma^2_{A.} = 1.0$
11	U1	NO	12.028	7.147	4.008	0.168	7.371	0.158	
12	U1	EX	12.030	7.657	3.999	0.560	7.771	0.558	$\rho = 0.6667$
13	NO	U1	12.051	9.555	4.010	0.086	9.851	0.078	Alpha = 0.6652
14	NO	NO	12.029	9.492	4.009	0.169	9.931	0.158	$E(MS_A) = 12.0$
15	NO	EX	12.076	10.484	3.999	0.560	10.331	0.558	
16	EX	U1	12.005	23.121	4.012	0.086	22.651	0.078	$E(MS_e) = 4.0$
17	EX	NO	11.883	22.484	4.010	0.170	22.731	0.158	
18	EX	EX	11.913	23.485	3.998	0.558	23.131	0.558	
19	U1	U1	6.875	2.910	3.997	0.088	2.853	0.078	$\sigma^2_{A.} = 0.36$
20	U1	NO	6.869	2.886	4.012	0.171	2.933	0.158	
21	U1	EX	6.868	3.180	4.003	0.526	3.333	0.558	$\rho = 0.4186$
22	NO	U1	6.911	3.161	3.997	0.088	3.184	0.078	Alpha = 0.4177
23	NO	NO	6.883	3.363	4.012	0.173	3.264	0.158	$E(MS_A) = 6.880$
24	NO	EX	6.858	3.629	4.003	0.526	3.664	0.558	
25	EX	U1	6.862	4.534	3.996	0.087	4.843	0.078	$E(MS_e) = 4.0$
26	EX	NO	6.865	4.627	4.012	0.173	4.923	0.158	
27	EX	EX	6.854	4.995	4.004	0.526	5.323	0.558	

$$(5.6)\begin{cases} \text{(a)} \quad \mathrm{Var}\,(MS_A) = \left[\frac{2}{I-1} + \frac{1}{I}\{\rho^2\gamma_A + (1-\rho)^2\,\gamma_e/J\}\right](J\sigma_A^2 + \sigma_e^2)^2 \\ \text{(b)} \quad \mathrm{Var}\,(MS_e) = \left[\frac{2}{(I-1)(J-1)} + \frac{\gamma_e}{IJ}\right]\sigma_e^4 \end{cases}$$

TABLE 5.12

Comparisons of Observed Means and Standard Errors of Reliability Estimates Under the Congeneric True Score Model With the Values Obtainable From Various Formulas, N = 2000, I = 30, J = 8

Exp. No.	Dis. Tr.	Dis. Er.	Observed $\hat{\rho}$ Mean (1)	Observed $\hat{\rho}$ S.E. (2)	S.E. by formulas (5.3) (3)	S.E. by formulas (5.1)-(b) (4)	S.E. by formulas (5.10) (5)	Parameters and expected values under ANOVA (6)
01	UI	UI	0.882	0.028	0.038	0.036	0.030	$\sigma_A^2 = 4.0$
02	UI	NO	0.883	0.028	0.038	0.036	0.030	
03	UI	EX	0.884	0.033	0.038	0.036	0.030	$\rho = 0.8889$
04	NO	UI	0.879	0.036	0.038	0.036	0.036	Alpha = 0.8870
05	NO	NO	0.879	0.036	0.038	0.036	0.036	$E(\hat{\rho}) = 0.8807$
06	NO	EX	0.880	0.040	0.038	0.036	0.036	
07	EX	UI	0.862	0.062	0.038	0.036	0.057	
08	EX	NO	0.862	0.061	0.038	0.036	0.057	
09	EX	EX	0.863	0.064	0.038	0.036	0.057	
10	UI	UI	0.645	0.091	0.101	0.108	0.098	$\sigma_A^2 = 1.0$
11	UI	NO	0.649	0.091	0.101	0.108	0.098	
12	UI	EX	0.651	0.098	0.101	0.108	0.100	$\rho = 0.6667$
13	NO	UI	0.643	0.106	0.101	0.108	0.108	Alpha = 0.6652
14	NO	NO	0.643	0.105	0.101	0.108	0.108	
15	NO	EX	0.646	0.114	0.101	0.108	0.109	$E(\hat{\rho}) = 0.6420$
16	EX	UI	0.613	0.155	0.101	0.108	0.146	
17	EX	NO	0.611	0.156	0.101	0.108	0.147	
18	EX	EX	0.615	0.157	0.101	0.108	0.148	
19	UI	UI	0.379	0.175	0.151	0.188	0.180	$\sigma_A^2 = 0.36$
20	UI	NO	0.379	0.172	0.151	0.188	0.182	
21	UI	EX	0.383	0.179	0.151	0.188	0.189	$\rho = 0.4186$
22	NO	UI	0.378	0.186	0.151	0.188	0.186	Alpha = 0.4177
23	NO	NO	0.373	0.189	0.151	0.188	0.188	
24	NO	EX	0.378	0.186	0.151	0.188	0.195	$E(\hat{\rho}) = 0.3755$
25	EX	UI	0.359	0.213	0.151	0.188	0.216	
26	EX	NO	0.356	0.219	0.151	0.188	0.217	
27	EX	EX	0.360	0.217	0.151	0.188	0.224	

$$
(5.1) \quad \begin{cases} \text{(a)} \quad E(\hat{\rho}) = 1 - (1-\rho)\,\dfrac{I-1}{I-3} \\[2ex] \text{(b)} \quad \mathrm{Var}\,(\hat{\rho}) = (1-\rho)^2 \;\dfrac{2(I-1)(\nu+I-3)}{(J-1)(I-3)^2(I-5)} \end{cases}
$$

$$
(5.3) \qquad \mathrm{Var}\,(\hat{\rho}) = \frac{(1-\rho^2)^2}{I}
$$

$$
(5.10) \qquad \mathrm{Var}\,(\hat{\rho}) \simeq (1-\rho)^2 \left[\frac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \frac{\gamma_y}{I} \right]
$$

reliability under the congeneric model. However, as can be seen in Figure 5.1, the shapes of the distributions are almost the same as expected from (2.17), namely under the ANOVA model and normal distribution theory. Therefore similar conclusions as cited in Section 5.1.5 may be obtained from the observation of Table 5.13, namely the real significance levels of the F-test, or the lower and upper 5% critial points of $\hat{\rho}$ are almost the same as the values under the ANOVA model.

To make the comparisons between the ANOVA model and the congeneric model, and to separate the effects of non-homogeneous true scores variances from the effects of non-homogeneous error variances, further experiments were performed under the same conditions as the ANOVA model cases except that the true score variances were allowed to differ. However, there is some possibility of the existence of interaction effects between the effects of violating the two homogeneity assumptions, although each case was found to be quite robust against the violations.

To investigate this problem, 15 additional experiments were performed employing three sets of $\underline{\lambda}$'s as before and five sets of non-homogeneous error variances used in Section 5.2.0, namely EV3, EV4, EV5, EV6, and EV7. The results are summarized in Tables 5.14, 5.15, and 5.16. When the entries of these tables are compared with the corresponding values of Tables 5.5, 5.6, and 5.7, little difference is noted between the two sets of values suggesting non-existence of such interaction effects. For example, experiment 3 of Table 5.5 gives the observed variance of MS_A as 88.222, while the corresponding value under the congeneric model is given in experiment 1 of Table 5.14 as 89.270. Therefore, it may be concluded that, the effect of non-

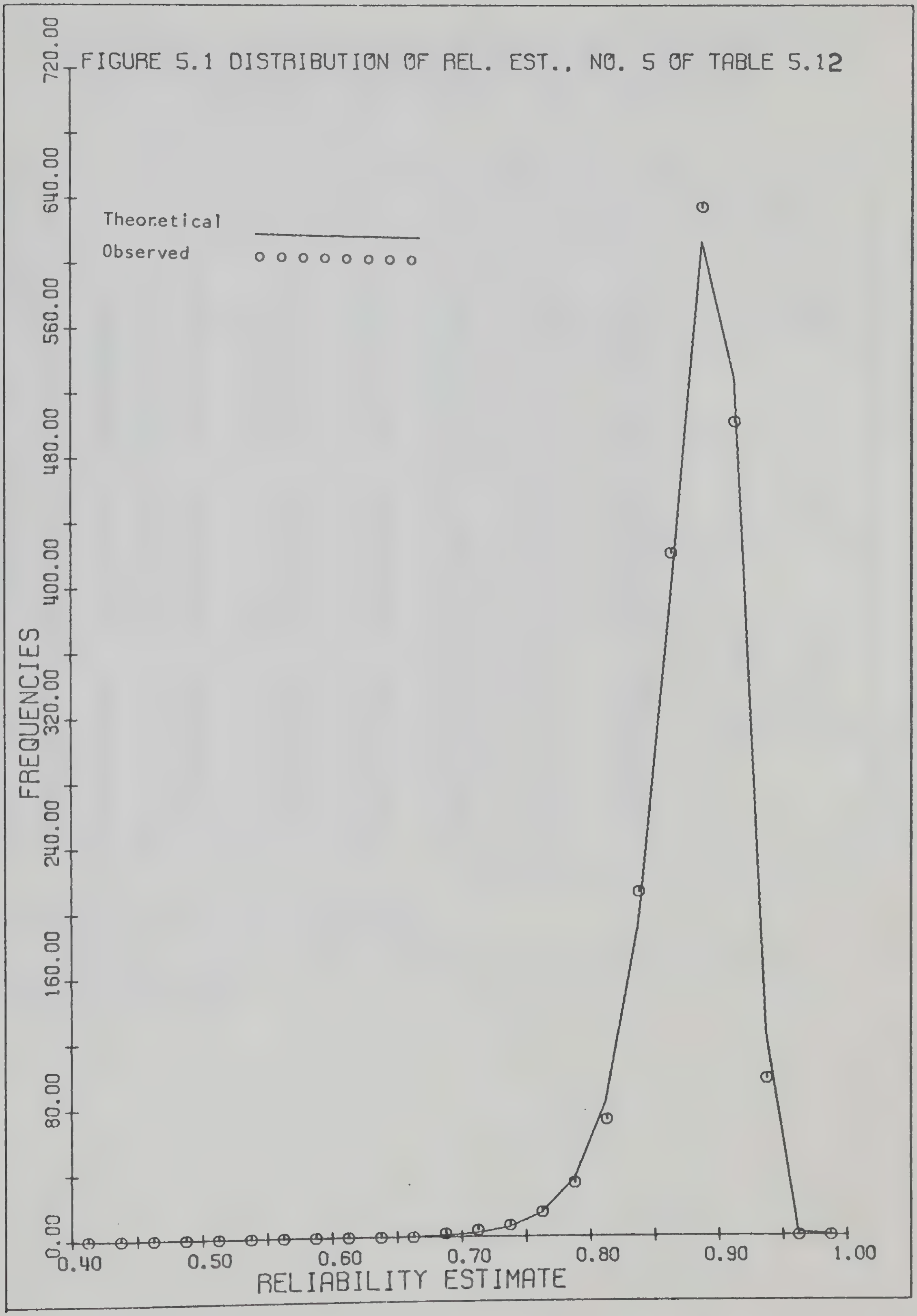


FIGURE 5.1 DISTRIBUTION OF REL. EST., NO. 5 OF TABLE 5.12
Theoretical
Observed
FREQUENCIES
0.00
80.00
160.00
240.00
320.00
400.00
480.00
560.00
640.00
720.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

TABLE 5.13

Comparisons of Observed Lower and Upper 5% Critical Points of Reliability Estimates Under the Congeneric True Score Model with the Values Obtainable Under the ANOVA and Normal Theory, and Real Type One Error of F-Test When the Nominal Value is 5%, N = 2000, I = 30, J = 8

Exp. No.	Dis. Tr.	Er.	True Sig. (%) Lower (1)	Upper (2)	Observed C.P.[1] Lower (3)	Upper (4)	Theoretical C.P.[2] Lower (5)	Upper (6)	Parameters (7)
01	U1	U1	1.90	1.35	0.834	0.917	0.814	0.927	$\sigma^2_{A.} = 4.0$
02	U1	NO	1.85	1.65	0.835	0.918	0.814	0.927	
03	U1	EX	2.75	4.05	0.826	0.925	0.814	0.927	$\rho = 0.8889$
04	NO	U1	5.15	2.90	0.813	0.924	0.814	0.927	Alpha = 0.887
05	NO	NO	4.60	3.95	0.816	0.925	0.814	0.927	
06	NO	EX	5.65	6.95	0.810	0.930	0.814	0.927	
07	EX	U1	17.95	9.64	0.745	0.938	0.814	0.927	
08	EX	NO	17.85	10.45	0.743	0.940	0.814	0.927	
09	EX	EX	18.35	11.70	0.750	0.942	0.814	0.927	
10	U1	U1	3.35	2.80	0.475	0.768	0.442	0.781	$\sigma^2_{A.} = 1.0$
11	U1	NO	3.05	3.35	0.482	0.771	0.442	0.781	
12	U1	EX	3.75	4.85	0.470	0.781	0.442	0.781	$\rho = 0.6667$
13	NO	U1	4.60	4.70	0.448	0.778	0.442	0.781	Alpha = 0.6652
14	NO	NO	4.60	4.80	0.447	0.780	0.442	0.781	
15	NO	EX	6.05	7.10	0.428	0.791	0.442	0.781	
16	EX	U1	11.70	10.45	0.325	0.812	0.442	0.781	
17	EX	NO	12.05	11.05	0.328	0.812	0.442	0.781	
18	EX	EX	12.80	11.85	0.319	0.820	0.442	0.781	
19	U1	U1	3.85	4.85	0.058	0.618	0.027	0.618	$\sigma^2_{A.} = 0.36$
20	U1	NO	3.65	4.35	0.068	0.610	0.027	0.618	
21	U1	EX	4.65	4.60	0.040	0.614	0.027	0.618	$\rho = 0.4186$
22	NO	U1	4.80	4.55	0.030	0.614	0.027	0.618	Alpha = 0.4177
23	NO	NO	5.25	4.60	0.020	0.616	0.027	0.618	
24	NO	EX	5.00	4.65	0.027	0.617	0.027	0.618	
25	EX	U1	7.15	6.95	-0.027	0.635	0.027	0.618	
26	EX	NO	8.40	7.20	-0.048	0.645	0.027	0.618	
27	EX	EX	7.40	8.40	-0.052	0.648	0.027	0.618	

[1]Observed lower and upper 5% critical points.

[2]Theoretical lower and upper 5% critical points under ANOVA with normal distribution of true and error scores.

TABLE 5.14

Comparisons of Observed Means and Variances of MS's Under Congeneric True Scores, Non-Homogeneous Error Variances and the Normal Distributions With the Values Obtainable Under ANOVA Model, N = 2000, I = 30, J = 8

Exp. No.	Er. Type	Observed MS_A Mean (1)	Observed MS_A Var. (2)	Observed MS_e Mean (3)	Observed MS_e Var. (4)	$E(MS_A)$ (5)	Var. by (5.6) MS_A (6)	Var. by (5.6) MS_e (7)	Parameters (8)
01	EV3	36.101	89.270	3.832	0.224	35.750	88.142	0.139	$\sigma^2_{A.} = 4.0$
02	EV4	35.611	83.103	3.630	0.239	35.562	87.220	0.125	
03	EV5	36.308	87.363	4.321	0.441	36.250	90.625	0.178	
04	EV6	38.742	103.316	6.472	0.708	38.375	101.561	0.400	
05	EV7	36.915	96.880	4.819	0.618	36.750	93.142	0.222	
06	EV3	11.795	9.654	3.758	0.241	11.750	9.522	0.139	$\sigma^2_{A.} = 1.0$
07	EV4	11.633	9.566	3.574	0.223	11.562	9.220	0.125	
08	EV5	12.249	10.121	4.693	0.432	12.250	10.349	0.178	
09	EV6	14.477	15.206	6.376	0.702	14.375	14.251	0.400	
10	EV7	12.740	11.321	4.775	0.609	12.750	11.211	0.222	
11	EV3	6.700	3.203	3.768	0.221	6.630	3.032	0.139	$\sigma^2_{A.} = 0.36$
12	EV4	6.430	2.779	3.564	0.232	6.442	2.862	0.125	
13	EV5	7.120	3.491	4.258	0.430	7.130	3.506	0.178	
14	EV6	9.306	5.915	6.394	0.771	9.255	5.907	0.400	
15	EV7	7.593	3.782	4.753	0.620	7.630	4.015	0.222	

$E(MS_e) = \sigma^2_{e.}$ = 3.750 (EV3)
3.563 (EV4)
4.250 (EV5)
6.375 (EV6)
4.750 (EV7)

$$(5.6)\quad \begin{cases} \text{(a)} \quad \mathrm{Var}\ (MS_A) = [\frac{2}{I-1} + \frac{1}{I}\{\rho^2\gamma_A + (1-\rho)^2\gamma_e/J\}](J\sigma^2_A + \sigma^2_e)^2 \\ \text{(b)} \quad \mathrm{Var}\ (MS_e) = [\frac{2}{(I-1)(J-1)} + \frac{\gamma_e}{IJ}]\ \sigma^4_e \end{cases}$$

TABLE 5.15

Comparisons of the Observed Means and Standard Error of Reliability Estimates Under Congeneric True Score, Non-Homogeneous Error Variances and Normal Distributions With the Values Obtainable From Formulas (5.1), (5.3), and (5.10), N = 2000, I = 30, J = 8

Exp. No.	Er. Var.	Rel. (1)	Alpha (2)	$E(\hat{\rho})$ by (5.1)-(a) (3)	Observed $\hat{\rho}$ Mean (4)	S.E. (5)	S.E. by (5.3) (6)	(5.10) (7)	Parameter (8)
01	EV3	0.895	0.893	0.887	0.886	0.036	0.036	0.034	$\sigma^2_{A.} = 4.0$
02	EV4	0.900	0.898	0.892	0.891	0.034	0.035	0.032	
03	EV5	0.883	0.881	0.874	0.873	0.041	0.040	0.038	
04	EV6	0.834	0.832	0.822	0.821	0.055	0.056	0.054	
05	EV7	0.871	0.869	0.861	0.860	0.048	0.044	0.042	
06	EV3	0.681	0.679	0.657	0.658	0.108	0.098	0.103	$\sigma^2_{A.} = 1.0$
07	EV4	0.692	0.690	0.669	0.670	0.102	0.095	0.100	
08	EV5	0.653	0.652	0.627	0.628	0.114	0.105	0.112	
09	EV6	0.557	0.555	0.524	0.528	0.140	0.126	0.143	
10	EV7	0.628	0.626	0.600	0.601	0.119	0.111	0.121	
11	EV3	0.434	0.434	0.393	0.397	0.180	0.148	0.183	$\sigma^2_{A.} = 0.36$
12	EV4	0.447	0.446	0.406	0.409	0.173	0.146	0.179	
13	EV5	0.404	0.403	0.360	0.365	0.181	0.153	0.193	
14	EV6	0.311	0.311	0.260	0.268	0.207	0.165	0.223	
15	EV7	0.378	0.377	0.331	0.337	0.190	0.157	0.201	

$$(5.1)\quad \begin{cases} \text{(a)} & E(\hat{\rho}) = 1 - (1-\rho)\dfrac{I-1}{I-3} \\ \text{(b)} & \mathrm{Var}(\hat{\rho}) = (1-\rho)^2 \dfrac{2(I-1)(\nu+I-3)}{(J-1)(I-3)^2(I-5)} \end{cases}$$

$$(5.3)\quad \mathrm{Var}(\hat{\rho}) = \frac{(1-\rho^2)^2}{I}$$

$$(5.10)\quad \mathrm{Var}(\hat{\rho}) \simeq (1-\rho)^2 \left[\frac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \frac{\gamma_Y}{I}\right]$$

TABLE 5.16

Comparisons of Observed Lower and Upper 5% Critical Points Under Congeneric True Scores, Non-Homogeneous Error Score Variance, and Normal Distributions With the Values Obtainable Under the ANOVA and Normal Theory, and Real Type One Errors of F-Test When the Nominal Value is 5%, N = 2000, I = 30, J = 8

Exp. No.	Er. Var.	Rel.	$E(\hat{\rho})$ by (5.1)-(b)	Observed C.P.[1]		Theoretical C.P.[2]		Real Sig. (%)		Parameter
				Lower	Upper	Lower	Upper	Lower	Upper	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
01	EV3	0.895	0.887	0.821	0.930	0.824	0.931	5.90	4.50	$\sigma^2_{A.} = 4.0$
02	EV4	0.900	0.892	0.829	0.934	0.832	0.934	5.90	4.80	
03	EV5	0.883	0.874	0.795	0.924	0.804	0.923	6.50	5.60	
04	EV6	0.834	0.822	0.720	0.892	0.722	0.891	5.25	5.25	
05	EV7	0.871	0.861	0.774	0.917	0.784	0.915	6.50	6.20	
06	EV3	0.681	0.657	0.448	0.792	0.466	0.791	5.80	5.10	$\sigma^2_{A.} = 1.0$
07	EV4	0.692	0.669	0.485	0.802	0.484	0.798	5.00	5.60	
08	EV5	0.653	0.627	0.417	0.773	0.419	0.772	5.15	5.15	
09	EV6	0.557	0.524	0.271	0.713	0.258	0.709	4.40	5.95	
10	EV7	0.628	0.600	0.377	0.757	0.376	0.755	4.90	5.20	
11	EV3	0.434	0.393	0.059	0.631	0.053	0.629	4.80	5.35	$\sigma^2_{A.} = 0.36$
12	EV4	0.447	0.406	0.096	0.627	0.074	0.637	4.70	3.90	
13	EV5	0.404	0.360	0.022	0.602	0.002	0.609	4.10	4.05	
14	EV6	0.311	0.260	-0.130	0.545	-0.153	0.548	4.30	4.70	
15	EV7	0.378	0.331	-0.035	0.582	-0.042	0.591	4.75	4.15	

[1]Observed lower and upper 5% critical points of $\hat{\rho}$.

[2]Theoretical lower and upper 5% critical points of $\hat{\rho}$ under ANOVA model.

homogeneous true score variance, or violation of ETEM assumption will affect the sampling distribution of the reliability estimates given by (2.17) very little as long as the degree of non-homogeneity is within a moderate range as with the $\underline{\lambda}$'s used in these experiments.

5.3.3 Distributions Under the Multi-Factor Model

Classically, the assumption of a one-factor true score has been referred to as one which produces 'unit rank correlation matrix' (e.g., Kuder and Richardson, 1937), but as seen in Chapter Two, the unifactorness of true scores is inherent to the ANOVA linear model and its more general form such as the ETEM or congeneric models. Under these models, it is implicitly assumed that the test measures only one trait, and therefore, the true score can have only one factor structure. However, in real test score data, it is seldom possible to separate measurement of one trait from others. The psychological or achievement tests usually measure more than one trait at a time, and it is sometimes unrealistic to assume that only one factor exists and to regard all other factors as error. This fact has been well demonstrated by the rejection by many researchers of Spearman's so-called g-factor theory in modern factor analysis. Thus violation of unifactor true score assumption may not be considered simply as an exceptional case; this may be rather a common case for real data.

In Chapter Two, the multi-factor test model has been introduced as a generalization of the congeneric test model by expanding the linear model of ANOVA step by step to a factor analytic model. However, since most of the test theories are based on the unifactor true score assumption, no reliability theory has ever been developed under this

model. Therefore the multi-factor model has been referred to as an assumption violating case of the classical model rather than a separate model in its own right. Following this traditional line, in this study, the reliability distribution under the multi-factor model is treated as an assumption violating case of the ANOVA model as are other models examined in the previous sections.

Since the Alpha coefficient is a measure of the first factor concentration (Cronbach, 1951), the coefficient is expected to be much lower than the reliability coefficient if second or higher factor is not negligible. Therefore, it is hardly expected that the sampling distribution of Alpha coefficient, as a substitute for the reliability estimate, is robust against the violation of the unifactor assumption.

To support this conjecture, a number of sampling experiments were performed and the results are compared with those obtainable under ANOVA model. The parameters σ_A^2 and σ_e^2 are not defined under this model as with the congeneric model case, but the average of true and error score variance may be used to determine the effectiveness of the ANOVA model under the multi-factor model, namely,

$$(5.15) \qquad \sigma_{A.}^2 = (\underline{1}'\,\underline{\Lambda}\,\underline{\Lambda}'\,\underline{1})/J^2 = (\sum_j \sum_{j'} \sum_r \lambda_{jr}\,\lambda_{j'r})/J^2 ,$$

and $\sigma_{e.}^2$ as in the previous section.

If these parameters are used in place of σ_A^2 and σ_e^2, most of the formulas introduced in Section 5.1.0 can be used directly and the robustness of the ANOVA model under multi-factor true score cases can be examined empirically by the simulation techniques.

Using the following two $\underline{\Lambda}$ matrices, an error score standard deviation matrix $\underline{\Psi}$,, and three types of true and error score distributions, altogether 18 experiments were performed under the multi-factor true score model with $N = 2000$, $I = 30$, $J = 6$,

$$\underline{\Lambda}_1 = \begin{bmatrix} 0.887 & 0.302 \\ 0.410 & 0.663 \\ 0.242 & 0.735 \\ 0.369 & 0.816 \\ 0.417 & 0.557 \\ 0.669 & 0.482 \end{bmatrix}, \qquad \underline{\Lambda}_2 = \frac{1}{2}\underline{\Lambda}_1 = \begin{bmatrix} 0.4435 & 0.1510 \\ 0.2050 & 0.3315 \\ 0.1210 & 0.3675 \\ 0.1845 & 0.4080 \\ 0.2085 & 0.2785 \\ 0.3345 & 0.2410 \end{bmatrix}$$

$$\underline{\Psi} = \begin{bmatrix} 0.34942 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.62636 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.63341 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.44495 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.71823 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.56579 \end{bmatrix}.$$

Table 5.17 compares the observed means and variances of the MS's under the multi-factor model with the values obtainable from formula (5.6) treating the model as an ANOVA model. It is noted that rather close agreement exists between the means of the observed MS_A and $E(MS_A)$ given in colums (1) and (7), but the agreement is rather poor between the means of observed MS_e and $E(MS_e)$ given in columns (3) and (7) indicating the effect of the violation of unifactor true score assumption. Even for the normal true and error score distributions, the difference between the two values are too big to be explained as sampling fluctuation. For example, experiment 5 gives the mean of MS_e as 0.413 while the theoretical value of $E(MS_e) = 0.3249$ if the ANOVA model and $\sigma^2_{e.}$ are used. This implies that the $E(MS_e)$ undervalues the real expected value of MS_e. When the variances of MS's, in colums (2) and (4), are compared with the values obtainable from formula (5.6)

TABLE 5.17

Comparisons of Observed Means and Variances of MS's Under the Multi-Factor True Score Model and Various Combinations of True and Error Score Distributions With the Values Obtainable Under ANOVA Model by Formula (5.6), N = 2000, I = 30, J = 6

Exp. No.	Dis Tr.	Er.	Observed MS_A Mean (1)	Observed MS_A Var. (2)	Observed MS_e Mean (3)	Observed MS_e Var. (4)	Var. by (5.6) MS_A (5)	Var. by (5.6) MS_e (6)	Parameters and E(MS) (7)
01	UI	UI	3.916	0.761	0.436	0.0017	0.542	0.0008	$\sigma^2_{A.} = 0.6001$
02	UI	NO	3.915	0.800	0.414	0.0026	0.544	0.0015	
03	UI	EX	3.902	0.814	0.414	0.0067	0.548	0.0050	$\sigma^2_{e.} = 0.3249$
04	NO	UI	3.919	1.087	0.413	0.0019	1.062	0.0008	
05	NO	NO	3.919	1.066	0.413	0.0027	1.063	0.0015	$E(MS_A) = 3.9235$
06	NO	EX	3.913	1.036	0.413	0.0066	1.066	0.0050	
07	EX	UI	3.898	2.341	0.414	0.0027	3.655	0.0008	$E(MS_e) = 0.3249$
08	EX	NO	3.895	2.278	0.414	0.0036	3.656	0.0015	
09	EX	EX	3.887	2.309	0.413	0.0073	3.659	0.0050	
10	UI	UI	1.233	0.086	0.345	0.0010	0.070	0.0008	$\sigma^2_{A.} = 0.1500$
11	UI	NO	1.220	0.089	0.346	0.0019	0.071	0.0015	
12	UI	EX	1.218	0.089	0.346	0.0058	0.075	0.0050	$\sigma^2_{e.} = 0.3249$
13	NO	UI	1.231	0.103	0.348	0.0010	0.103	0.0008	
14	NO	NO	1.223	0.103	0.346	0.0019	0.103	0.0015	$E(MS_A) = 1.225$
15	NO	EX	1.225	0.109	0.348	0.0059	0.107	0.0050	
16	EX	UI	1.220	0.181	0.346	0.0011	0.265	0.0008	$E(MS_e) = 0.3249$
17	EX	NO	1.216	0.174	0.347	0.0020	0.266	0.0015	
18	EX	EX	1.242	0.188	0.346	0.0063	0.269	0.0050	

$$
(5.6)\quad \begin{cases} \text{(a)} \quad \mathrm{Var}\,(MS_A) = \left[\frac{2}{I-1} + \frac{1}{I}\{\rho^2\gamma_A + (1-\rho)^2\,\gamma_e/J\}\right](J\sigma_A^2 + \sigma_e^2)^2 \\ \text{(b)} \quad \mathrm{Var}\,(MS_e) = \left[\frac{2}{(I-1)(J-1)} + \frac{\gamma_e}{IJ}\right]\sigma_e^4 \end{cases}
$$

given in columns (5) and (6), rather poor agreement is noticed, suggesting inapplicability of the formula.

Table 5.18 gives means and standard errors of reliability estimates and compares them with the values obtainable from formulas (5.3), (5.1)-(b), and (5.10). It is observed that, for all experiments, the mean of $\hat{\rho}$ is much lower then population Alpha or $E(\hat{\rho})$ under ANOVA and normal theory, indicating the effect of the multi-factor true score structure. This result is probably due to the fact that the Alpha coefficient measures mainly the variance due to the first factor, and thus underestimates the true score variance and overestimates the error score variance, and at the same time shifting the distribution of reliability estimates considerably to the left as shown in Figure 5.2. Although formula (5.10) seems still to be the best among the three, the fit is very poor suggesting inapplicability of most of the formulas derived under the ANOVA model and normal theory for multi-factor true score test.

Discrepancies between observed and theoretical distributions based on ANOVA model are clearly seen when the real significance level of F-test is compared with the nominal value of 5%, as summarized in Table 5.19. The real significance level for the lower tail range from 14.40% to 27.50% for $\underline{\Lambda}_1$ and 5.80% to 12.30% for $\underline{\Lambda}_2$, clearly indicating the inapplicability of the conventional F-test to multi-factor tests. For the upper tail, the true significance levels are in general lower than the nominal value, but the results are not predictable. For example, experiment 2 gives a value as low as 0.45%, while experiment 18 gives one as high as 8.10%. All of these results

TABLE 5.18

Comparisons of Observed Means and Standard Errors of Reliability Estimates Under the Multi-Factor True Score Model and Various Combinations of True and Error Score Distributions With the Values Obtainable From Formulas (5.3), (5.1)-(b) and (5.10), N = 2000, I = 30, J = 6

Ex. No.	Dis. Tr.	Er.	Observed $\hat{\rho}$ Mean (1)	S.E. (2)	Calculated S.E. by (5.3) (3)	(5.1)-(b) (4)	(5.10) (5)	Parameters and $E(\hat{\rho})$, ANOVA and Normal (6)
01	UI	UI	0.888	0.032	0.029	0.027	0.023	$\sigma^2_{A.} = 0.6001$
02	UI	NO	0.888	0.033	0.029	0.027	0.023	
03	UI	EX	0.887	0.039	0.029	0.027	0.023	$\sigma^2_{e.} = 0.3249$
04	NO	UI	0.886	0.037	0.029	0.027	0.027	
05	NO	NO	0.886	0.038	0.029	0.027	0.027	
06	NO	EX	0.887	0.038	0.029	0.027	0.027	$\rho = 0.9172$
07	EX	UI	0.879	0.042	0.029	0.027	0.044	Alpha = 0.8943
08	EX	NO	0.879	0.047	0.029	0.027	0.044	$E(\hat{\rho}) = 0.9111$
09	EX	EX	0.880	0.049	0.029	0.027	0.044	
10	UI	UI	0.702	0.085	0.084	0.088	0.078	$\sigma^2_{A.} = 0.1500$
11	UI	NO	0.698	0.088	0.084	0.088	0.079	
12	UI	EX	0.700	0.097	0.084	0.088	0.080	$\sigma^2_{e.} = 0.3249$
13	NO	UI	0.695	0.093	0.084	0.088	0.088	
14	NO	NO	0.696	0.096	0.084	0.088	0.088	$\rho = 0.7348$
15	NO	EX	0.696	0.106	0.084	0.088	0.089	Alpha = 0.7164
16	EX	UI	0.682	0.115	0.084	0.088	0.124	$E(\hat{\rho})$ - 0.7151
17	EX	NO	0.682	0.118	0.084	0.088	0.124	
18	EX	EX	0.691	0.120	0.084	0.088	0.124	

(5.1)
$$\begin{cases} \text{(a)} \quad E(\hat{\rho}) = 1 - (1-\rho)\,\dfrac{I-1}{I-3} \\[2ex] \text{(b)} \quad \operatorname{Var}(\hat{\rho}) = (1-\rho)^2\,\dfrac{2(I-1)(\nu+I-3)}{(J-1)(I-3)^2(I-5)} \end{cases}$$

(5.3)
$$\operatorname{Var}(\hat{\rho}) = \frac{(1-\rho^2)^2}{I}$$

(5.10)
$$\operatorname{Var}(\hat{\rho}) = (1-\rho)^2\left[\frac{2(I-1)(IJ-J-2)}{(J-1)(I-3)^2(I-5)} + \frac{\gamma_y}{I}\right]$$

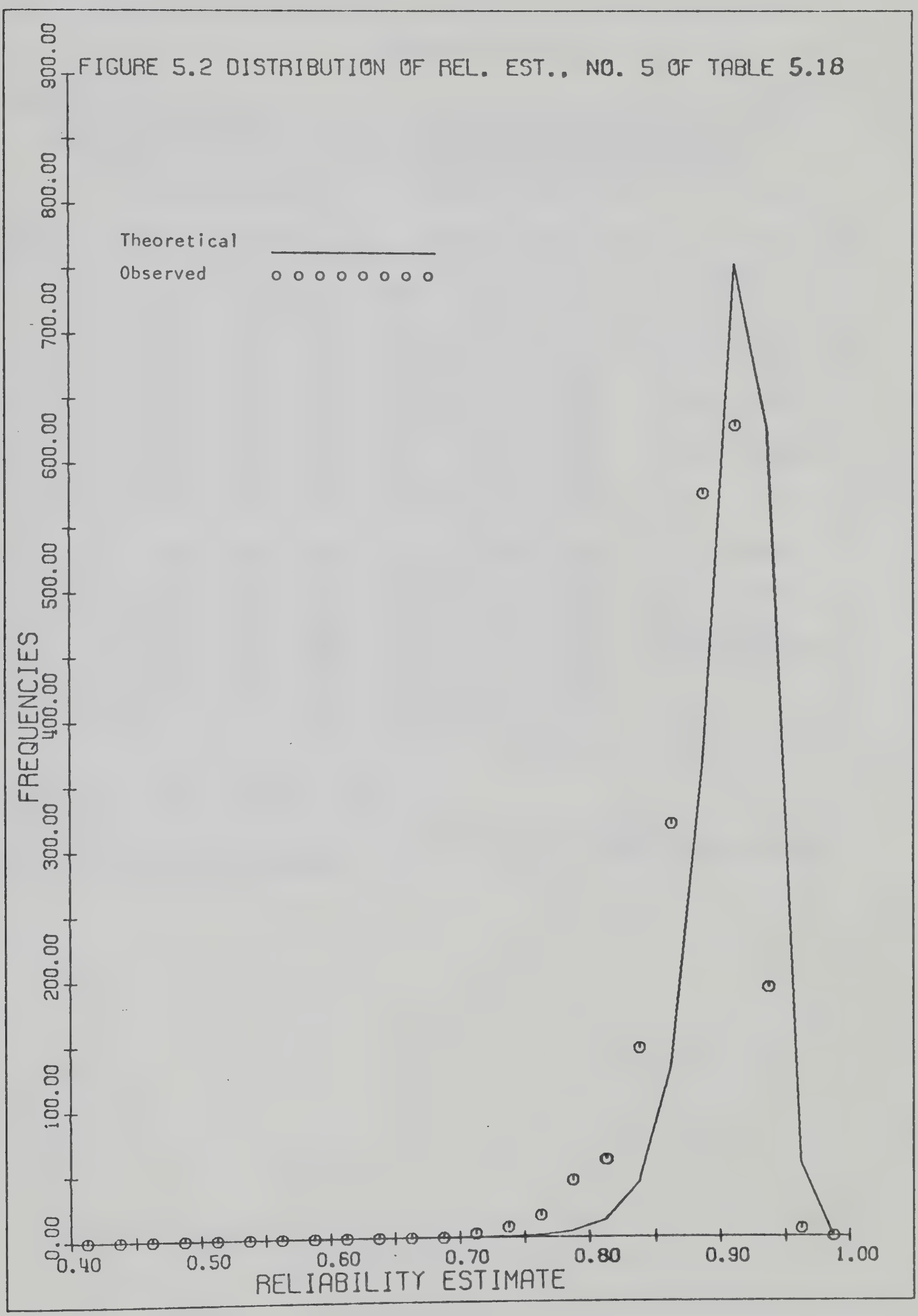


FIGURE 5.2 DISTRIBUTION OF REL. EST., NO. 5 OF TABLE 5.18
Theoretical
Observed
FREQUENCIES
0.00
100.00
200.00
300.00
400.00
500.00
600.00
700.00
800.00
900.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

TABLE 5.19

Comparisons of Observed Lower and Upper 5% Critical Points of Reliability Estimates and Real Type One Errors of F-Test When the Nominal Value is 5% Under the Multi-Factor True Score Model and Various Combinations of True and Error Score Distributions With the Values Obtainable Under the ANOVA Model and Normal Theory, N = 2000, I = 30, J = 6

Exp. No.	Dis. Tr.	Dis. Er.	Real Sig. (%) Lower (1)	Real Sig. (%) Upper (2)	Observed C.P.[1] Lower (3)	Observed C.P.[1] Upper (4)	Theoretical C.P.[2] Lower (5)	Theoretical C.P.[2] Upper (6)	Parameters (7)
01	UI	UI	15.35	0.50	0.813	0.929	0.860	0.946	$\sigma^2_{A.} = 0.6001$
02	UI	NO	14.40	0.45	0.828	0.930	0.860	0.946	
03	UI	EX	19.50	1.45	0.817	0.934	0.860	0.946	$\sigma^2_{e.} = 0.3249$
04	NO	UI	19.80	0.40	0.820	0.931	0.860	0.946	
05	NO	NO	18.90	0.55	0.815	0.932	0.860	0.946	$\rho = 0.9172$
06	NO	EX	20.90	1.45	0.819	0.936	0.860	0.946	Alpha = 0.8943
07	EX	UI	27.50	1.25	0.814	0.939	0.860	0.946	
08	EX	NO	26.45	1.55	0.797	0.936	0.860	0.946	
09	EX	EX	26.80	2.40	0.786	0.939	0.860	0.946	
10	UI	UI	5.80	2.25	0.541	0.813	0.552	0.828	$\sigma^2_{A.} = 0.1500$
11	UI	NO	6.85	2.20	0.530	0.815	0.552	0.828	
12	UI	EX	7.95	4.05	0.522	0.825	0.552	0.828	$\sigma^2_{e.} = 0.3249$
13	NO	UI	7.55	2.40	0.524	0.814	0.552	0.828	
14	NO	NO	8.20	2.95	0.508	0.814	0.552	0.828	$\rho = 0.7348$
15	NO	EX	8.45	4.70	0.500	0.826	0.552	0.828	Alpha = 0.7164
16	EX	UI	12.06	5.00	0.475	0.828	0.552	0.828	
17	EX	NO	12.30	4.85	0.450	0.828	0.552	0.828	
18	EX	EX	11.56	8.10	0.466	0.845	0.552	0.828	

[1]Observed lower and upper 5% critical points of $\hat{\rho}$.

[2]Theoretical lower and upper 5% critical points of $\hat{\rho}$ under the ANOVA model with normal distribution of true and error scores.

strongly suggest that ANOVA model and normal theory are not robust against the violation of the assumption of unifactor true score.

5.3.4 Conclusions for the Effects of Non-ETEM Model

Based on the above discussions, the following conclusions are tentatively made.

(a) Formula (5.6) may be used for the congeneric test case if $\sigma^2_{A.}$ and $\sigma^2_{e.}$ are used in place of σ^2_A and σ^2_e of ANOVA model. However, this formula is valueless for the case of multi-factor true score model.

(b) The non-homogeneity of true score variance has little effect on the distribution, although the ETEM assumption is violated if the violation is moderate. The conclusions obtained in Section 5.1.5 may be generalized to the congeneric true score cases with moderate violation of ETEM assumption.

(c) The effects of violation of the unifactor true score assumption are the most critical. If this assumption is violated, the formulas derived under the ANOVA model cannot be applied directly even with a normal true score distribution.

(d) The F-test based on (2.17) may be used for the congeneric model if the true score distribution is approximately normal as in the ANOVA model case and the homogeneity of true score variance is satisfied approximately, but it would be misleading in multi-factor model cases. This is especially true for inferences based on lower portions or high reliability case. As previous sections showed, these effects diminish with the lower values of reliability.

Findings in this section are based on rather limited combinations of possible parameters and distributions of true and

error scores, and therefore, generalization must be made with care.

5.4.0 Summary

Sampling distributions of reliability estimates for the continuous part-test cases are investigated under the ANOVA, ETEM, and congeneric and multi-factor true score models with various combinations of true and error scores distributions by analytical and computer simulation methods. Tukey's results for the calculation of the variance of variance estimate under an ANOVA model were applied to test theory to obtain an approximate formula for standard error of reliability estimates when the distributions of true and error scores are not necessary normal.

To investigate sampling distributions of reliability estimates based on formula (2.13) under these models and distributional assumptions not necessarily normal, to see robustness of the ANOVA model and normal theory represented by the formula (2.17), and to evaluate the new formula for the standard error of reliability estimates, altogether 156 experiments were performed by REL01, each requiring approximately 6 minutes of computer C.P.U. time. From the experiments, the following conclusions may be obtained.

(a) The equation (2.17) obtained under the ANOVA model and normal theory is quite robust against the violation of the following assumptions if the reliability estimate is based on (2.13), i.e., the estimation formula for Alpha coefficient:

i) Normality of error score distributions.
ii) Homogeneity of error score variances.
iii) Homogeneity of true score variances, if violation is moderate.

But the ANOVA model and normal theory is not robust against violation of the following assumptions.

i) Unifactorness of true score dispersion matrix.

ii) Normality of true score distributions.

The effects of the violation of these last two assumptions will decrease as the values of reliability decrease.

(b) For the F-test based on the equation (2.17), the multi-factor true score model increases Type one error for the lower tail and decreases it for the upper tail by shifting the distributions of reliability estimates leftward substantially, when second or higher factors of the true score dispersion matrix cannot be ignored.

(c) The effects of non-normal true score distributions depend on the magnitude of their kurtosis. For negative kurtosis, Type one errors for both tails are less than the nominal value, while for positive kurtosis, they are greater than the nominal value. The greater the absolute value of kurtosis, the greater is the discrepancy from the nominal value.

(d) If true scores are distributed as normal, the ANOVA, ETEM, and congeneric models give almost identical distributions of reliability estimates with moderate departures from homogeneity assumptions of error and/or true score variances.

(e) The new standard error formula (5.10) is superior to the traditional formulas (5.1) or (5.3), if the true scores are not distributed as normal.

CHAPTER SIX

RESULTS FOR BINARY ITEM TEST SCORE CASES

This chapter presents the results of computer simulated experiments for the binary item test score cases. Section 6.1 deals with the overall factors which might affect the distribution of reliability estimates. In Section 6.2, the effects of non-normal error distributions are investigated with normal latent score distributions and homogeneous biserial correlations. Section 6.3 deals with the cases of non-normal latent scores with homogeneous biserial correlations and normal error scores, while Section 6.4 deals with non-normal latent scores and non-homogeneous biserial correlations. For all cases, both homogeneous and non-homogeneous item difficulty parameters are employed to determine the effects of non-homogeneous difficulty parameters.

6.1 Factors Related to Binary Item Test Scores Distribution

As discussed in Chapter Three, for a composite test consisting of J binary items as its part-tests, direct decomposition of observed score x_{ij}, which takes the value unity for a correct response and zero otherwise, into two independent parts, namely true and error scores, is impossible. Thus the linear model equation (3.3) can only be applied to an intervening variable or 'response strength variable' y_{ij} which is a hypothetical continuous variable.

Under the normal ogive model, it was possible to evaluate test parameters such as the variance σ_x^2, reliability ρ, and KR20 by means of numerical methods if the item parameters, such as difficulty

parameters $\{\pi_j\}$ and biserial correlations $\{\lambda_j\}$, are specified. Unfortunately, however, the computational formulas given in Chapter Three are valid if and only if the normal ogive model is valid, namely, if the $\{f_i\}$ and $\{\varepsilon_{ij}\}$ are independently and identically distributed as $N(0,1)$ as discussed in Section 4.6 of Chapter Four. Thus the non-normal distributions of these two types of random variables would affect not only the sampling distribution of reliability estimates, but also the population test parameters.

Furthermore, for the continuous part score case, the fixed constant for each part, β_j, indicates the relative difficulty level of each part-test, but these parameters do not enter any formula for reliability or any other test parameters, and are independent of the sampling distribution of reliability estimates. Therefore it was not necessary to consider the effects of $\{\beta_j\}$ on the distribution of reliability estimates. For the binary item case, however, the item difficulty parameters, the analogue of β_j for the continuous case, enter the formula (3.15) through threshold constants and consequently affect such test score parameters, as the mean, variance, reliability and KR20. Furthermore, as shown in Section 3.4 of Chapter Three, the ETEM assumption is satisfied if and only if the items are all homogeneous, namely they have equal difficulty and biserial correlation parameters. Therefore, if the difficulty parameters are not homogeneous, it is expected that the KR20 will be lower than the reliability and subsequently the sampling distribution of reliability estimates may differ from that of the homogeneous case, though there is some indication that the effects are not great (Nitko and Feldt, 1969).

As a result, for the binary item cases, the following factors must be taken into account for a study of sampling distributions of reliability estimates:

(a) The effect of non-normal distributions of $\{f_i\}$ and $\{\varepsilon_{ij}\}$, i.e., the effect of the violation of the normal ogive model.

(b) Homogeneity of item difficulty parameters and biserial correlations, i.e., the effect of the violation of the ETEM assumption.

Obviously it is impossible to investigate the sampling distributions of reliability estimates under all possible combinations of the above factors and all possible sets of parameters by computer simulation techniques. In this chapter, to conserve the overall computer time, the experiments and investigations are limited to only three distributions for $\{f_i\}$ and $\{\varepsilon_{ij}\}$, namely uniform (U1), normal (NO), and exponential (EX); four sets of difficulty parameters, two of which are non-homogeneous, and six sets of biserial correlations, three sets of which are non-homogeneous. The parameter sets used for the experiments are given in Tables 6.1 and 6.2

TABLE 6.1

Item Difficulty Parameters

Notation	Homogeneity	Item Number 1	2	3	4	5	6	7	8	9	Mean	Var.
D1	Homo.	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.0
D2	Non-Homo.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.5	0.0667
D3	Homo.	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.0
D4	Non-Homo.	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.7	0.0167

TABLE 6.2

Item Biserial Correlations

Notation	Homogeneity	Item Number 1	2	3	4	5	6	7	8	9	Mean	Var.
B1	Homo.	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.0
B2	Non-Homo.	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.7	0.0167
B3	Homo.	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.0
B4	Non-Homo.	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.6	0.0167
B5	Homo.	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.0
B6	Non-Homo.	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.4	0.0167

With these distribution-parameter combinations, altogether 216 experiments ($3 \times 3 \times 4 \times 6$) are possible. However, previous results indicated that the error distribution has little effect on the distribution of reliability estimates, and the same tendency may be expected for the binary item cases. Since this was the case, as will be seen in the following section, only the first step of the investigation will involve the case of non-normal error distributions. Thus the total number of experiments run were 96, resulting in a saving of computer time.

6.2 The Effects of Non-Normal Error Distribution and Non-Homogeneous Item Difficulty Parameters

In order to separate possible effects of non-normal latent score distribution and non-homogeneous biserial correlations such as B2, B4, and B6, from those of non-homogeneous item difficulty parameters or non-normal errors, which are of major interest in this section, three homogeneous biserial correlation sets B1, B3, and B5, normal distribution of latent variables, four sets of difficulty parameters, three types

of error distributions are used, i.e., altogether 36 (3×4×3) experiments with N = 1000, I = 30 and J = 9 are performed.

Table 6.3 presents population parameters calculated from the formulas given in Chapter Three and the results obtained from the parallel form method, with sample size 30030. Comparisons of data in Table 6.3 indicate:

(a) Calculated test parameters based on formulas given in Chapter Three agree reasonably well with the results obtained by computer simulation, thus partially validating the computer simulation method. For example, experiment 2 was performed with normal error score distribution, and satisfies the normal ogive model. It gives the test score mean, variance, reliability, and KR20 as 4.491, 8.094, 0.813, and 0.812, while the theoretical values based on the normal ogive model are 4.5, 8.118, 0.813, and 0.813 respectively.

(b) For normal latent score distributions, the observed test score means given in column (5) seem to depend only on the average of the item difficulty parameters as expected, and are affected neither by non-homogeneous difficulty parameters nor non-normal error score distributions. For example, the values of experiments 1- 6 inclusive in column (5) are almost identical to theoretical value 4.5, although experiments 1, 3, 4, 6 have non-normal error score distributions, and experiments 4, 5, and 6 have non-homogeneous difficulty parameters.

(c) The non-homogeneous difficulty parameter sets, D2 and D4, (e.g., experiments 4, 5, 6, and 10, 11, 12) result in lower test score variance, reliability, and KR20 when compared with the same average level of difficulty, but homogeneous, namely D1, and D3

TABLE 6.3

Comparisons of Calculated Test Parameters Under the Normal Ogive Model With Empirical Values Based on the Parallel Form Method, Normal Latent Scores, and Homogeneous Biserial Correlations, NI = 30030, J = 9

Exp. No.	Err. Dis.	Bis.	Dif.	Theoretical (N.O.) Mean (1)	Var. (2)	Rel. (3)	KR20 (4)	Observed by P.F.M. Mean (5)	Var. (6)	Rel. (7)	KR20 (8)
01	U1	B1	D1	4.5	8.118	0.813	0.813	4.494	8.044	0.811	0.810
02	NO	B1	D1	4.5	8.118	0.813	0.813	4.491	8.094	0.813	0.812
03	EX	B1	D1	4.5	8.118	0.813	0.813	4.484	8.103	0.812	0.813
04	U1	B1	D2	4.5	4.979	0.769	0.752	4.496	4.941	0.768	0.750
05	NO	B1	D2	4.5	4.979	0.769	0.752	4.495	5.012	0.772	0.754
06	EX	B1	D2	4.5	4.979	0.769	0.752	4.506	4.968	0.768	0.751
07	U1	B1	D3	6.3	6.589	0.802	0.802	6.289	6.493	0.802	0.802
08	NO	B1	D3	6.3	6.589	0.802	0.802	6.299	6.621	0.804	0.804
09	EX	B1	D3	6.3	6.589	0.802	0.802	6.281	6.622	0.803	0.803
10	U1	B1	D4	6.3	5.671	0.788	0.780	6.269	5.714	0.789	0.780
11	NO	B1	D4	6.3	5.671	0.788	0.780	6.296	5.632	0.785	0.777
12	EX	B1	D4	6.3	5.671	0.788	0.780	6.283	5.689	0.787	0.780
13	U1	B3	D1	4.5	6.470	0.734	0.734	4.514	6.447	0.732	0.732
14	NO	B3	D1	4.5	6.470	0.734	0.734	4.498	6.460	0.733	0.733
15	EX	B3	D1	4.5	6.470	0.734	0.734	4.494	6.487	0.735	0.735
16	U1	B3	D2	4.5	4.092	0.684	0.671	4.480	4.129	0.687	0.675
17	NO	B3	D2	4.5	4.092	0.684	0.671	4.487	4.053	0.678	0.675
18	EX	B3	D2	4.5	4.092	0.684	0.671	4.499	4.079	0.681	0.670
19	U1	B3	D3	6.3	5.230	0.718	0.718	6.286	5.259	0.719	0.719
20	NO	B3	D3	6.3	5.230	0.718	0.718	6.276	5.215	0.715	0.715
21	EX	B3	D3	6.3	5.230	0.718	0.718	6.267	5.276	0.719	0.719
22	U1	B3	D4	6.3	4.559	0.702	0.696	6.298	4.573	0.703	0.697
23	NO	B3	D4	6.3	4.559	0.702	0.696	6.295	4.546	0.701	0.694
24	EX	B3	D4	6.3	4.559	0.702	0.696	6.283	4.557	0.700	0.694
25	U1	B5	D1	4.5	4.091	0.506	0.506	4.483	4.063	0.502	0.502
26	NO	B5	D1	4.5	4.091	0.506	0.506	4.498	4.104	0.509	0.508
27	EX	B5	D1	4.5	4.091	0.506	0.506	4.500	4.141	0.513	0.514
28	U1	B5	D2	4.5	2.735	0.452	0.446	4.497	2.696	0.442	0.437
29	NO	B5	D2	4.5	2.735	0.452	0.446	4.495	2.727	0.453	0.450
30	EX	B5	D2	4.5	2.735	0.452	0.446	4.490	2.776	0.461	0.455
31	U1	B5	D3	6.3	3.317	0.484	0.484	6.304	3.347	0.490	0.490
32	NO	B5	D3	6.3	3.317	0.484	0.484	6.295	3.243	0.469	0.469
33	EX	B5	D3	6.3	3.317	0.484	0.484	6.292	3.344	0.486	0.488
34	U1	B5	D4	6.3	2.956	0.466	0.463	6.297	2.959	0.465	0.461
35	NO	B5	D4	6.3	2.956	0.466	0.463	6.292	2.980	0.469	0.466
36	EX	B5	D4	6.3	2.956	0.466	0.463	6.289	2.955	0.467	0.462

respectively (e.g., experiments 1, 2, 3, and 7, 8, 9). For non-homogeneous difficulty, i.e., D2 and D4, the KR20 coefficients are lower than the reliability as expected since the ETEM assumption is not satisfied. For example, experiment 12, with non-homogeneous difficulty set of D4, gives reliability and KR20 as 0.787 and 0.780 respectively.

(d) For homogeneous item difficulty, the higher the item difficulty is above the ideal 0.5 level, the lower the test variance, reliability and KR20. The same trends are observed for difficulty lower than 0.5 level, though the results are not reported in this paper since almost exactly the same results as high difficulty cases are obtained for lower difficulty cases except for test means, i.e., the test parameters, except for the test means, are highest when the item difficulty parameters are all equal to 0.5 which is a well-known fact in test theory. For example, experiment 1 has homogeneous difficulty of 0.5 for all items and gives variance and reliability as 8.044 and 0.811 respectively, while experiment 7, which is comparable to experiment 1 except the higher difficulty of 0.7, gives 6.493, and 0.802 respectively. However, this conclusion would not apply in general to the non-homogeneous item difficulty cases, i.e., the non-homogeneous item difficulty effects interact with the effects of item difficulty level, and the results are not predictable, as it can be seen when the results of experiments 4, 5, and 6 are compared with those of experiments 10, 11, and 12.

(e) The non-normal distributions of error scores have very little effect on the test parameters. For example, experiment 12, which has an exponential error distribution, gives parameter values as 6.283, 5.689, 0.787, and 0.780 which can be compared reasonably well with

theoretical values given in columns (1)-(4) inclusive, namely 6.3, 5.671, 0.788, and 0.780 respectively. Alternatively, they can also be compared reasonably well with the corresponding values of experiment 11 which has a normal error distribution, namely 6.296, 5.632, 0.785, and 0.777.

Table 6.4 gives the means and standard error of reliability estimates over N = 1000 trials and compares them with theoretical values which can be obtained from continuous part scores under the ANOVA model and normal distributional theory, i.e., treating binary test scores $\{x_{ij}\}$ as if they were continuous part scores as in the previous chapter. From the table, it is noted that the observed means of reliability estimates given in column (2), which is based on estimation formula (2.13), compares fairly well with the theoretical values given in column (4) based on (5.1)-(a), the largest difference being only 0.018 (experiment 32) which is probably too small to be meaningful in test theory. The standard error obtained from formula (5.3) or (5.1)-(b), given in columns (5) and (6), also predict the observed standard errors given in column (3) reasonably well, although formula (5.3) seems to consistently underestimate the standard errors for lower reliability cases, namely the case of biserial correlation set B5. In general, formula (5.1)-(b) seems quite satisfactory, the largest difference between the theoretical and observed values being only 0.0111 (experiment 28). The sum of squares from the observed standard errors are 0.00101 and 0.00956 for formulas (5.1)-(b) and (5.3) respectively, suggesting the superiority of formulas (5.1)-(b) to (5.3). No attempts are made to use formula (5.10) since neither kurtosis formulas of test

TABLE 6.4

Comparisons of Observed Means and Standard Errors of Reliability Estimates Under Normal Latent Scores, Homogeneous Biserial Correlations With the Values Obtainable From ANOVA Model and Normal Theory, N = 1000, I = 30, J = 9

Exp. No.	Err. Dis.	Bis.	Dif.	Rel.* (1)	Observed $\hat{\rho}$ Mean (2)	Observed $\hat{\rho}$ S.E. (3)	$E(\hat{\rho})$ by (5.1)-(a) (4)	Expected S.E. by (5.3) (5)	Expected S.E. by (5.1)-(b) (6)
01	UI	B1	D1	0.811	0.800	0.0591	0.797	0.0626	0.0608
02	NO	B1	D1	0.813	0.802	0.0600	0.799	0.0618	0.0600
03	EX	B1	D1	0.812	0.804	0.0561	0.799	0.0621	0.0602
04	UI	B1	D2	0.768	0.736	0.0714	0.751	0.0749	0.0745
05	NO	B1	D2	0.769	0.742	0.0667	0.752	0.0745	0.0740
06	EX	B1	D2	0.768	0.738	0.0685	0.751	0.0748	0.0744
07	UI	B1	D3	0.802	0.787	0.0688	0.787	0.0653	0.0637
08	NO	B1	D3	0.802	0.789	0.0708	0.788	0.0651	0.0635
09	EX	B1	D3	0.803	0.789	0.0693	0.788	0.0649	0.0633
10	UI	B1	D4	0.789	0.765	0.0737	0.773	0.0689	0.0677
11	NO	B1	D4	0.788	0.760	0.0770	0.772	0.0693	0.0682
12	EX	B1	D4	0.787	0.764	0.0786	0.771	0.0695	0.0683
13	UI	B3	D1	0.732	0.717	0.0821	0.712	0.0847	0.0859
14	NO	B3	D1	0.734	0.718	0.0826	0.714	0.0843	0.0855
15	EX	B3	D1	0.735	0.722	0.0793	0.715	0.0840	0.0851
16	UI	B3	D2	0.687	0.656	0.0947	0.664	0.0964	0.1004
17	NO	B3	D2	0.684	0.646	0.1035	0.660	0.0972	0.1015
18	EX	B3	D2	0.681	0.650	0.1024	0.657	0.0979	0.1024
19	UI	B3	D3	0.719	0.701	0.0932	0.698	0.0881	0.0901
20	NO	B3	D3	0.718	0.694	0.0986	0.698	0.0883	0.0904
21	EX	B3	D3	0.719	0.700	0.0917	0.699	0.0881	0.0901
22	UI	B3	D4	0.703	0.676	0.0970	0.681	0.0923	0.0953
23	NO	B3	D4	0.702	0.672	0.0995	0.680	0.0927	0.0957
24	EX	B3	D4	0.700	0.673	0.0959	0.677	0.0932	0.0964
25	UI	B5	D1	0.502	0.469	0.1585	0.465	0.1367	0.1600
26	NO	B5	D1	0.506	0.474	0.1619	0.470	0.1358	0.1585
27	EX	B5	D1	0.513	0.482	0.1561	0.477	0.1345	0.1563
28	UI	B5	D2	0.442	0.401	0.1682	0.400	0.1470	0.1793
29	NO	B5	D2	0.452	0.407	0.1807	0.411	0.1453	0.1759
30	EX	B5	D2	0.461	0.415	0.1733	0.421	0.1438	0.1731
31	UI	B5	D3	0.490	0.452	0.1636	0.453	0.1387	0.1636
32	NO	B5	D3	0.484	0.428	0.1739	0.446	0.1398	0.1656
33	EX	B5	D3	0.486	0.452	0.1614	0.448	0.1394	0.1650
34	UI	B5	D4	0.465	0.425	0.1719	0.425	0.1431	0.1718
35	NO	B5	D4	0.466	0.428	0.1776	0.427	0.1429	0.1713
36	EX	B5	D4	0.467	0.425	0.1615	0.427	0.1428	0.1711

*Theoretical value if error scores are normal, otherwise the value was obtained by the parallel form method.

(5.1) (a) $E(\hat{\rho}) = 1 - (1-\rho)\frac{I-1}{I-3}$ (b) $\mathrm{Var}(\hat{\rho}) = (1-\rho)^2 \frac{2(I-1)(\nu+I-3)}{(J-1)(I-3)^2(I-5)}$

(5.3) $\mathrm{Var}(\hat{\rho}) = \frac{(1-\rho^2)^2}{I}$

score for binary item test nor any numerical means to evaluate the parameter are available at present.

Table 6.5 indicates the shapes of the lower and upper portions of the distributions of reliability estimates by giving the lower and upper 5% critical points of the distributions in columns (2) and (3), and by comparing them with those results obtainable theoretically from (2.17), given in columns (4) and (5). The results, in general, suggest that the theoretical values are very close to the observed values except for the upper tail portions for some experiments with high or medium reliability, i.e., with B1 and B2, and non-homogeneous difficulty set D2, namely experiments 4, 5, 6, 16, 17, and 18. For those experiments, the distributions are systematically shifted toward lower reliability primarily due to the fact that KR20 is substantially lower than reliability, because of extreme non-homogeneity of item difficulty parameter set D2. This is illustrated in Figure 6.1. Consequently the real Type one errors of the F-test for upper tails are much smaller than the nominal 5% level, some dropping as low as 1.1% level [column (7) of experiment 4]. The effect of non-homogeneous item difficulty parameters on the real significance level diminishes as the variation of item difficulty parameters decrease, as shown by experiments 10, 11, 12, 22, 23, and 24. The effect of item difficulty also diminishes with lower reliability level, and no meaningful differences are observed for low reliability cases illustrated by experiments 25-36 inclusive.

From the above observations, the following conclusions are tentatively made.

TABLE 6.5

Comparisons of Observed Lower and Upper 5% Critical Points Under Normal Latent Scores and Homogeneous Biserial Correlations With the Values Obtainable From the ANOVA Model and Normal Theory, and Real Type One Error of F-Test When Nominal Value is Fixed to the 5% Level, N = 1000, I = 30, J = 9

Exp. No.	Err. Dis.	Bis.	Dif.	Rel.[1]	Observed C.P.[2] Lower	Upper	Theoretical C.P.[3] Lower	Upper	Real Sig. (0/0) Lower	Upper
				(1)	(2)	(3)	(4)	(5)	(6)	(7)
01	U1	B1	D1	0.811	0.693	0.877	0.684	0.875	4.30	5.60
02	NO	B1	D1	0.813	0.698	0.882	0.688	0.877	3.80	7.40
03	EX	B1	D1	0.812	0.697	0.878	0.687	0.876	3.90	5.70
04	U1	B1	D2	0.768	0.604	0.828	0.612	0.847	5.60	1.10
05	NO	B1	D2	0.769	0.614	0.831	0.615	0.848	5.20	1.50
06	EX	B1	D2	0.768	0.605	0.825	0.613	0.847	5.70	1.60
07	U1	B1	D3	0.802	0.655	0.877	0.669	0.869	6.60	8.00
08	NO	B1	D3	0.802	0.650	0.879	0.670	0.870	6.70	6.80
09	EX	B1	D3	0.803	0.665	0.877	0.670	0.870	5.50	7.00
10	U1	B1	D4	0.789	9.629	0.857	0.648	0.861	7.30	4.00
11	NO	B1	D4	0.788	0.622	0.857	0.645	0.860	6.40	4.90
12	EX	B1	D4	0.787	0.623	0.858	0.644	0.860	6.70	4.60
13	U1	B3	D1	0.732	0.560	0.830	0.553	0.823	4.70	6.60
14	NO	B3	D1	0.734	0.572	0.828	0.555	0.824	3.90	5.60
15	EX	B3	D1	0.735	0.574	0.827	0.557	0.825	3.70	5.60
16	U1	B3	D2	0.687	0.487	0.783	0.478	0.794	4.30	3.40
17	NO	B3	D2	0.684	0.464	0.777	0.472	0.791	4.90	3.00
18	EX	B3	D2	0.681	0.465	0.776	0.467	0.790	5.10	2.40
19	U1	B3	D3	0.719	0.528	0.821	0.531	0.815	5.40	6.00
20	NO	B3	D3	0.718	0.505	0.825	0.530	0.814	6.20	7.90
21	EX	B3	D3	0.719	0.531	0.823	0.531	0.815	5.10	6.70
22	U1	B3	D4	0.703	0.507	0.808	0.504	0.804	4.90	5.60
23	NO	B3	D4	0.702	0.484	0.797	0.502	0.803	6.40	4.20
24	EX	B3	D4	0.700	0.482	0.801	0.498	0.802	6.40	4.60
25	U1	B5	D1	0.502	0.194	0.684	0.168	0.671	3.70	6.30
26	NO	B5	D1	0.506	0.189	0.678	0.176	0.674	4.70	5.30
27	EX	B5	D1	0.513	0.186	0.678	0.187	0.679	5.00	4.80
28	U1	B5	D2	0.442	0.093	0.621	0.067	0.632	4.20	4.00
29	NO	B5	D2	0.452	0.056	0.637	0.085	0.639	5.60	4.60
30	EX	B5	D2	0.461	0.069	0.637	0.099	0.644	5.80	4.20
31	U1	B5	D3	0.490	0.149	0.670	0.149	0.664	5.00	5.70
32	NO	B5	D3	0.484	0.106	0.663	0.138	0.660	5.20	6.70
33	EX	B5	D3	0.486	0.140	0.661	0.142	0.661	5.10	5.00
34	U1	B5	D4	0.465	0.097	0.647	0.106	0.647	5.10	5.00
35	NO	B5	D4	0.466	0.104	0.648	0.109	0.648	5.50	4.80
36	EX	B5	D4	0.467	0.126	0.650	0.110	0.649	4.40	5.20

[1] Theoretical value if error scores are normal, otherwise the value was obtained by the parallel form method.

[2] Observed lower and upper 5% critical points of the distribution of β.

[3] Theoretical lower and upper 5% critical points of the distribution of β under ANOVA and normal theory.

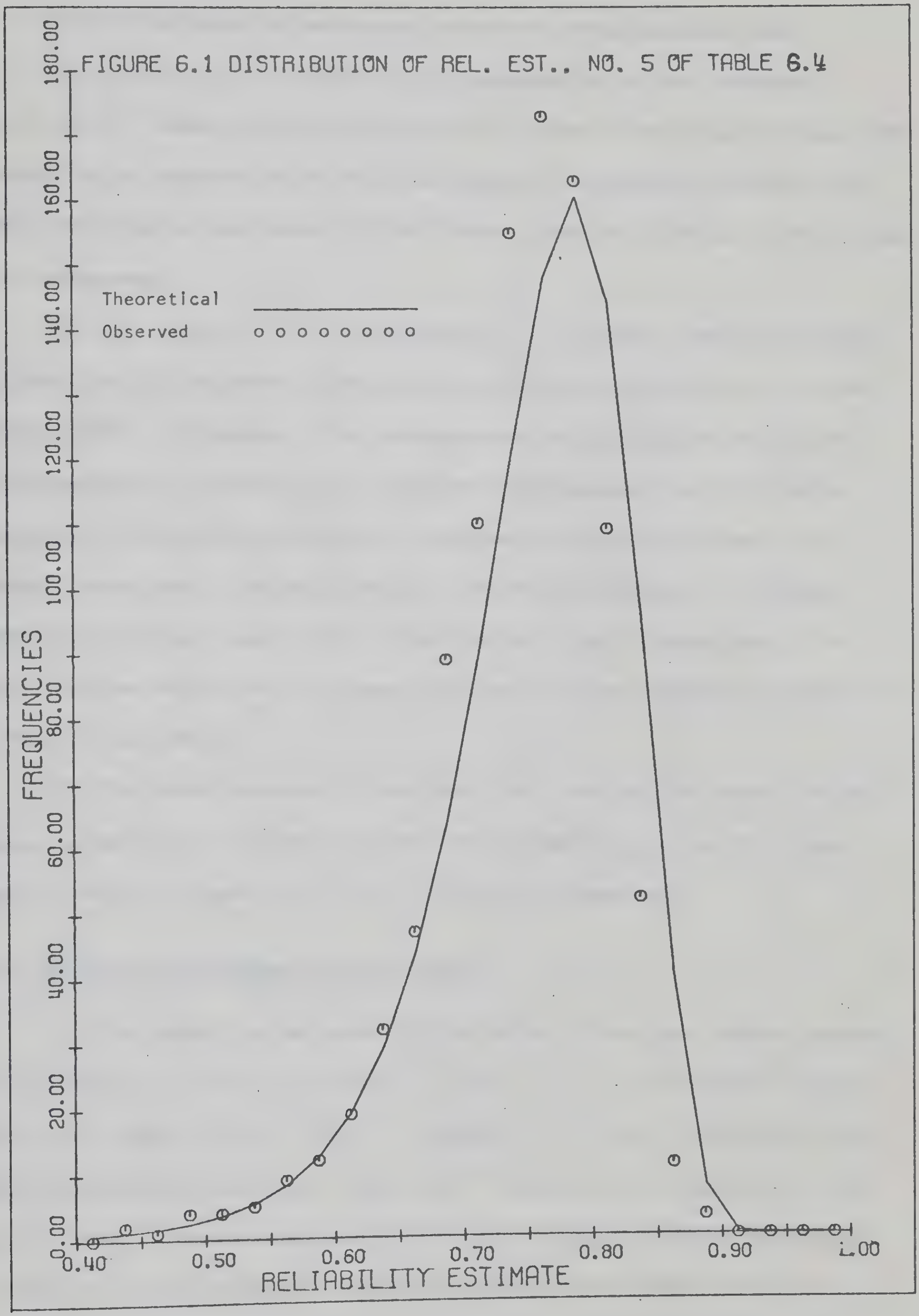


FIGURE 6.1 DISTRIBUTION OF REL. EST., NO. 5 OF TABLE 6.4
Theoretical
Observed
FREQUENCIES
0.00
20.00
40.00
60.00
80.00
100.00
120.00
140.00
160.00
180.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

(a) The effects of non-normal error distributions are small.

(b) Formula (5.1), both for the expected value and standard error of $\hat{\rho}$, seems quite satisfactory for binary item cases although the assumption of continuity of observed scores is violated, provided that the latent scores are normally distributed, and the biserial correlations are homogeneous.

(c) The item difficulty parameters $\{\pi_j\}$ affect the distribution systematically, contrary to the previous findings reported by Nitko and Feldt (1969). In general, the heterogeneity of difficulty shifts the distributions to the left, and the more heterogeneous the difficulty parameters, the more distortion is observed, and it also appears to cause a large shift leftward for high reliability cases. If F-tests based on (2.17) are used with a fixed nominal significance level, the real Type one errors for the upper tail portion are affected by the item difficulty parameters.

(d) The distributions of the lower tail portion for high or middle range reliability, or both tails for low reliability are quite stable against the heterogeneity of item difficulty parameters.

6.3 Effects of Non-Normal Latent Scores

The normal ogive model for the binary item test scores assumes the existence of latent variables or scores $\{f_i\}$ distributed independently and identically as $N(0,1)$. However, it is not conceivable that these assumptions are always satisfied. Therefore, the effects of non-normal latent distributions are one of the important factors which must be examined rather closely. For the continuous part score cases, it is known that the non-normal true scores affect the distribution of reliability estimates significantly, and inflate or deflate the real

Type one errors for the F-test. Thus it must be determined whether the same is true for the binary item test cases when the latent scores are not normal. Because it is known, from the experiments of the previous section, that the effects of non-normal errors are small, and to save computer time, experiments were performed using only normal error scores.

Using two kinds of non-normal latent score distributions, namely uniform (U1) and exponential (EX), four types of item difficulty sets and three kinds of homogeneous biserial correlation sets were selected. A total of 24 $(2 \times 4 \times 3)$ additional experiments were performed with $N = 1000$, $I = 30$, and $J = 9$. The results of these 24 experiments are summarized in Tables 6.6, 6.7, and 6.8, together with the results of 12 experiments of the previous section which uses normal latent and error scores, for the purposes of comparisons.

As in the previous section, the population parameters were first examined to determine the effects of non-normality of latent scores. From Table 6.6, it is clearly observed that the test means are almost identical for both methods, namely by theoretical calculations under the normal ogive model given in column (1) and by the parallel form method given in column (5), except for the exponential distributions which have non-zero skewness. Using the exponential distribution, the means are in general lower than the theoretical values suggesting the effects of skewness, since, unlike the variance, the means are in general more sensitive to non-zero skewness.

The effects of non-normal latent scores can be seen rather clearly when the observed variance, reliability and KR20, given in columns (6), (7), and (8), are examined. The values of variance,

TABLE 6.6

Comparisons of Calculated Test Parameters Under the Normal Ogive Model With Empirical Values Based on the Parallel Form Method, Normal Error Scores, Homogeneous Biserial Correlations, NI = 30030, J = 9

Exp. No.	Tr. Dis.	Bis	Dif.	Theoretical (N.O.) Mean (1)	Var. (2)	Rel. (3)	KR20 (4)	Observed by P.F.M. Mean (5)	Var. (6)	Rel. (7)	KR20 (8)
01	U1	B1	D1	4.5	8.118	0.813	0.813	4.504	9.034	0.845	0.845
02	NO	B1	D1	4.5	8.118	0.813	0.813	4.491	8.094	0.813	0.812
03	EX	B1	D1	4.5	8.118	0.813	0.813	4.156	6.997	0.766	0.765
04	U1	B1	D2	4.5	4.979	0.769	0.752	4.488	5.342	0.788	0.772
05	NO	B1	D2	4.5	4.979	0.769	0.752	4.495	5.012	0.772	0.754
06	EX	B1	D2	4.5	4.979	0.769	0.752	4.348	4.247	0.720	0.702
07	U1	B1	D3	6.3	6.589	0.802	0.802	6.235	7.081	0.820	0.821
08	NO	B1	D3	6.3	6.589	0.802	0.802	6.299	6.621	0.804	0.804
09	EX	B1	D3	6.3	6.589	0.802	0.802	6.209	4.589	0.653	0.653
10	U1	B1	D4	6.3	5.671	0.788	0.780	6.266	5.987	0.803	0.795
11	NO	B1	D4	6.3	5.671	0.788	0.780	6.296	5.632	0.785	0.777
12	EX	B1	D4	6.3	5.671	0.788	0.780	6.224	3.957	0.648	0.636
13	U1	B3	D1	4.5	6.470	0.734	0.734	4.498	7.025	0.764	0.765
14	NO	B3	D1	4.5	6.470	0.734	0.734	4.498	6.460	0.733	0.733
15	EX	B3	D1	4.5	6.470	0.734	0.734	4.299	5.610	0.675	0.675
16	U1	B3	D2	4.5	4.092	0.684	0.671	4.477	4.298	0.701	0.690
17	NO	B3	D2	4.5	4.092	0.684	0.671	4.487	4.053	0.678	0.666
18	EX	B3	D2	4.5	4.092	0.684	0.671	4.406	3.653	0.638	0.626
19	U1	B3	D3	6.3	5.230	0.718	0.718	6.232	5.568	0.738	0.738
20	NO	B3	D3	6.3	5.230	0.718	0.718	6.276	5.215	0.715	0.715
21	EX	B3	D3	6.3	5.230	0.718	0.718	6.228	3.822	0.561	0.560
22	U1	B3	D4	6.3	4.559	0.702	0.696	6.268	4.792	0.701	0.713
23	NO	B3	D4	6.3	4.559	0.702	0.696	6.295	4.546	0.701	0.694
24	EX	B3	D4	6.3	4.559	0.702	0.696	6.250	3.397	0.559	0.551
25	U1	B5	D1	4.5	4.091	0.506	0.506	4.512	4.199	0.522	0.522
26	NO	B5	D1	4.5	4.091	0.506	0.506	4.451	4.104	0.509	0.508
27	EX	B5	D1	4.5	4.091	0.506	0.506	4.451	3.834	0.464	0.465
28	U1	B5	D2	4.5	2.735	0.452	0.446	4.496	2.767	0.460	0.455
29	NO	B5	D2	4.5	2.735	0.452	0.446	4.495	2.747	0.453	0.450
30	EX	B5	D2	4.5	2.735	0.452	0.446	4.458	2.613	0.422	0.418
31	U1	B5	D3	6.3	3.317	0.484	0.484	6.279	3.384	0.492	0.494
32	NO	B5	D3	6.3	3.317	0.484	0.484	6.295	3.243	0.469	0.469
33	EX	B5	D3	6.3	3.317	0.484	0.484	6.272	2.828	0.366	0.369
34	U1	B5	D4	6.3	2.956	0.466	0.463	6.276	2.990	0.474	0.468
35	NO	B5	D4	6.3	2.956	0.466	0.463	6.292	2.980	0.469	0.466
36	EX	B5	D4	6.3	2.956	0.466	0.463	6.279	2.552	0.363	0.359

reliability and KR20 under uniform distributions are always higher than the corresponding values under normal distributions, while the values under exponential distributions are lower than the values under normal distributions. Thus the normal ogive model gives lower values for the uniform distribution cases than the real values and higher values for the exponential distribution. For example, with biserial correlation and difficulty parameters fixed to set B1 and D2, the theoretical values under the normal ogive model are 4.979, 0.769, and 0.752 for variance, reliability and KR20 respectively (experiment 5). The parallel form method under normal distribution gives 5.012, 0.772, and 0.754, closely approximating the theoretical values as expected. However, for uniform latent scores (experiment 4), the corresponding values are 5.342, 0.788, and 0.772, which are much higher than theoretical [given in columns (2), (3), and (4) of experiment 4] or observed values under normal distribution of latent scores [given in columns (6), (7), and (8) of experiment 5]. On the other hand, for exponential distributions (experiment 6), the observed values, i.e., 4.247, 0.720, and 0.702, are much less than the theoretical or observed values under normal distribution.

Therefore it may be concluded that the reliability parameters to be used for equation (2.17) must be ρ^*, the value obtained by the parallel form method, rather than ρ for non-normal latent score distribution cases, since these values are closer to the actual values than the theoretical values obtained under the normal distribution assumption of latent scores.

Table 6.7 presents the results for observed means and standard errors of reliability estimates using $N = 1000$, and compares them

TABLE 6.7

Comparisons of Observed Means and Standard Errors of Reliability Estimates Under Normal Errors and Homogeneous Biserial Correlations With the Values Obtainable From ANOVA Model and Normal Theory, N = 1000, I = 30, J = 9

Exp. No.	Tr. Dis.	Bis.	Dif.	Rel.* (1)	Observed $\hat{\rho}$ Mean (2)	Observed $\hat{\rho}$ S.E. (3)	$E(\hat{\rho})$ by (5.1)-(a) (4)	Expected S.E. by (5.3) (5)	Expected S.E. by (5.1)-(b) (6)
01	U1	B1	D1	0.845	0.838	0.0469	0.833	0.0522	0.0498
02	NO	B1	D1	0.813	0.802	0.0600	0.799	0.0618	0.0600
03	EX	B1	D1	0.766	0.751	0.0783	0.748	0.0755	0.0752
04	U1	B1	D2	0.788	0,762	0.0584	0.772	0.0693	0.0682
05	NO	B1	D2	0.769	0.742	0.0667	0.752	0.0745	0.0740
06	EX	B1	D2	0.720	0.680	0.1024	0.699	0.0880	0.0900
07	U1	B1	D3	0.820	0.811	0.0562	0.807	0.0597	0.0577
08	NO	B1	D3	0.702	0.789	0.0708	0.788	0.0651	0.0635
09	EX	B1	D3	0.653	0.636	0.1010	0.627	0.1048	0.1115
10	U1	B1	D4	0.803	0.786	0.0553	0.789	0.0647	0.0631
11	NO	B1	D4	0.788	0.760	0.0770	0.772	0.0693	0.0682
12	EX	B1	D4	0.648	0.620	0.0983	0.622	0.1059	0.1129
13	U1	B3	D1	0.764	0.754	0.0667	0.746	0.0761	0.0758
14	NO	B3	D1	0.734	0.718	0.0826	0.714	0.0843	0.0855
15	EX	B3	D1	0.675	0.654	0.1035	0.651	0.0993	0.1042
16	U1	B3	D2	0.701	0.675	0.0850	0.679	0.0928	0.0959
17	NO	B3	D2	0.684	0.646	0.1035	0.660	0.0972	0.1015
18	EX	B3	D2	0.638	0.598	0.1261	0.611	0.1082	0.1161
19	U1	B3	D3	0.738	0.723	0.0779	0.719	0.0831	0.0841
20	NO	B3	D3	0.718	0.694	0.0986	0.698	0.0883	0.0904
21	EX	B3	D3	0.561	0.537	0.1283	0.529	0.1250	0.1408
22	U1	B3	D4	0.719	0.697	0.0854	0.698	0.0882	0.0902
23	NO	B3	D4	0.702	0.672	0.0995	0.680	0.0927	0.0957
24	EX	B3	D4	0.559	0.528	0.1263	0.526	0.1256	0.1416
25	U1	B5	D1	0.522	0.494	0.1402	0.487	0.1327	0.1533
26	NO	B5	D1	0.506	0.474	0.1619	0.470	0.1358	0.1585
27	EX	B5	D1	0.464	0.429	0.1752	0.425	0.1432	0.1719
28	U1	B5	D2	0.460	0.419	0.1669	0.420	0.1440	0.1735
29	NO	B5	D2	0.452	0.407	0.1807	0.411	0.1453	0.1759
30	EX	B5	D2	0.422	0.374	0.1886	0.379	0.1501	0.1857
31	U1	B5	D3	0.492	0.461	0.1544	0.455	0.1383	0.1629
32	NO	B5	D3	0.848	0.428	0.1739	0.446	0.1398	0.1656
33	EX	B5	D3	0.366	0.330	0.1883	0.320	0.1581	0.2034
34	U1	B5	D4	0.474	0.437	0.1481	0.435	0.1416	0.1689
35	NO	B5	D4	0.466	0.428	0.1776	0.427	0.1429	0.1713
36	EX	B5	D4	0.636	0.322	0.1904	0.315	0.1586	0.2046

*Theoretical value if true scores are normal, otherwise the value was obtained by the parallel method.

(5.1) (a) $E(\hat{\rho}) = 1 - (1-\rho)\,\frac{I-1}{I-3}$ (b) $\mathrm{Var}\,(\hat{\rho}) = (1-\rho)^2\,\frac{2(I-1)(\nu+I-3)}{(I-1)(I-3)^2(I-5)}$

(5.3) $\mathrm{Var}\,(\hat{\rho}) = \frac{(1-\rho^2)^2}{I}$

with the calculated values based on formulas (5.1), and (5.3). It is noted that $E(\hat{\rho})$ of (5.1)-(a) given in column (4) predicts very well the observed means of reliability estimates given in column (2) regardless of the distributions of latent scores, the largest discrepancy being only 0.019 (experiment 6), suggesting robustness of the estimation formula (2.13) as far as point estimations are concerned. The observed standard error of estimation given in column (3) suggests that the uniform distributions of latent scores produces smaller standard errors while the exponential gives larger standard errors than under the normal distributions for high reliability cases. Formula (5.3) or (5.1)-(b) predicts the standard errors of reliability estimates reasonably well, though (5.1)-(b) seems better than (5.3).

Table 6.8 summarizes the shapes of the distributions of $\hat{\rho}$ at the tall portions by comparing lower and upper 5% critical points given in columns (2) and (3) with theoretical values given in columns (4) and (5). From the table, it may be concluded that the effects of item difficulty parameters as noted in the previous section can be generalized to non-normal latent score cases. On the other hand, from the observations of the real Type one errors, the effect of non-normal latent score distributions are not so obvious. The Type one errors are fluctuating substantially, but with no clear sign of systematic inflation or deflation of Type one errors due to non-normal distribution of latent scores, unlike the case of continuous part scores discussed in Chapter Five. This suggests robustness of the ANOVA model and normal distributional theory for the case of binary item tests.

From the above observations, the following conclusions were tentatively made.

TABLE 6.8

Comparisons of Observed Lower and Upper 5% Critical Points Under Normal Error Scores and Homogeneous Biserial Correlations With the Values Obtainable From ANOVA Model and Normal Theory, and Real Type One Error of F-Test When Nominal Value is Fixed to the 5% Level, N = 1000, I = 30, J = 9

Exp. No.	Tr. Dis.	Bis.	Dif.	Rel.[1]	Observed C.P.[2] Lower	Observed C.P.[2] Upper	Theoretical C.P.[3] Lower	Theoretical C.P.[3] Upper	Real Sig. (0/0) Lower	Real Sig. (0/0) Upper
				(1)	(2)	(3)	(4)	(5)	(6)	(7)
01	U1	B1	D1	0.845	0.749	0.901	0.741	0.898	3.80	6.30
02	NO	B1	D1	0.813	0.698	0.882	0.688	0.877	3.80	7.40
03	EX	B1	D1	0.766	0.611	0.851	0.609	0.846	4.60	6.70
04	U1	B1	D2	0.788	0.654	0.835	0.645	0.860	3.80	1.50
05	NO	B1	D2	0.769	0.614	0.831	0.615	0.848	5.20	1.50
06	EX	B1	D2	0.720	0.500	0.805	0.532	0.815	7.50	3.20
07	U1	B1	D3	0.820	0.716	0.887	0.700	0.882	3.20	6.70
08	NO	B1	D3	0.802	0.650	0.879	0.670	0.870	6.70	6.80
09	EX	B1	D3	0.653	0.446	0.775	0.420	0.771	3.70	5.70
10	U1	B1	D4	0.803	0.682	0.863	0.672	0.870	4.00	2.90
11	NO	B1	D4	0.788	0.622	0.857	0.645	0.860	6.40	4.90
12	EX	B1	D4	0.648	0.222	0.749	0.413	0.768	3.20	2.10
13	U1	B3	D1	0.764	0.631	0.844	0.605	0.844	3.00	5.10
14	NO	B3	D1	0.734	0.572	0.828	0.555	0.824	3.90	5.60
15	EX	B3	D1	0.675	0.450	0.790	0.458	0.786	5.40	5.50
16	U1	B3	D2	0.701	0.528	0.786	0.501	0.803	3.60	1.90
17	NO	B3	D2	0.684	0.464	0.777	0.472	0.791	4.90	3.00
18	EX	B3	D2	0.638	0.366	0.754	0.396	0.761	7.20	4.00
19	U1	B3	D3	0.738	0.582	0.829	0.562	0.827	3.80	5.60
20	NO	B3	D3	0.718	0.505	0.825	0.530	0.814	6.20	7.90
21	EX	B3	D3	0.561	0.290	0.703	0.267	0.711	4.20	3.80
22	U1	B3	D4	0.719	0.530	0.806	0.531	0.815	5.10	3.00
23	NO	B3	D4	0.702	0.484	0.797	0.502	0.803	6.40	4.20
24	EX	B3	D4	0.559	0.305	0.697	0.263	0.709	2.90	3.60
25	U1	B5	D1	0.522	0.243	0.683	0.203	0.685	3.60	4.70
26	NO	B5	D1	0.506	0.189	0.678	0.176	0.674	4.70	5.30
27	EX	B5	D1	0.464	0.115	0.653	0.106	0.647	4.80	5.70
28	U1	B5	D2	0.460	0.105	0.641	0.098	0.644	4.70	4.80
29	NO	B5	D2	0.452	0.056	0.637	0.085	0.639	5.60	4.60
30	EX	B5	D2	0.422	0.021	0.616	C.034	0.619	5.50	4.70
31	U1	B5	D3	0.492	0.166	0.669	0.152	0.665	4.50	5.30
32	NO	B5	D3	0.484	0.106	0.663	0.138	0.660	5.20	6.70
33	EX	B5	D3	0.366	-0.024	0.580	0.058	0.582	3.90	4.90
34	U1	B5	D4	0.474	0.179	0.643	0.121	0.653	3.00	4.00
35	NO	B5	D4	0.466	0.104	0.648	0.109	0.648	5.50	4.80
36	EX	B5	D4	0.363	-0.002	0.573	-0.064	0.580	3.20	4.00

[1]Theoretical values if true scores are normal, otherwise the values obtained by the parallel form method.

[2]Observed lower and upper 5% critical points of the distribution of $\hat{\rho}$.

[3]Theoretical lower and upper 5% critical points of the distribution of $\hat{\rho}$ under ANOVA and normal theory.

(a) The non-normal latent score distributions affect the test parameters such as variance, reliability and KR20, and with lesser degree the mean, if the distribution is skewed. The normal ogive model provides smaller values than the actual values of the variance, reliability and KR20 if the latent scores are distributed as uniform, and the opposite is true for exponential distribution.

(b) Formulas (5.1) and (5.3) are quite robust against the violation of assumptions of normality for the binary item score cases.

(c) The effects of item difficulty parameters are the same as observed in the previous section.

(d) The non-normal latent scores do not systematically inflate or deflate real Type one errors for the F-test. The F-test seems quite robust against the violation of distributional assumptions, if difficulty parameters are homogeneous.

6.4 Effects of Non-Homogeneous Biserial Correlations

For the previous two sections, the biserial correlations were limited to homogeneous cases, namely B1, B3, and B5. In this section, three non-homogeneous biserial correlation sets, B2, B4, and B6 are used to investigate the effects of such non-homogeneity. Since it is known that for the continuous part-test score cases the non-homogeneity of true score variance, which corresponds to the square of biserial correlation for the binary item case under the congeneric true score model, does not affect the sampling distribution of the reliability estimates if the non-homogeneity is moderate, and it is of interest to know whether the same conclusion can be made for the binary item cases.

Employing three kinds of latent score distributions, U1, NO, and EX, and four sets of difficulty parameters, as in the previous section, and three sets of non-homogeneous biserial correlation sets, altogether 36 $(3 \times 4 \times 3)$ additional experiments were performed with $N = 1000$, $I = 30$, and $J = 9$. The results are summarized in Tables 6.9, 6.10, and 6.11.

If the test parameters estimated by the parallel form method in Table 6.9 are compared with the corresponding entries of Table 6.6, the latter table using the same parameter distribution combinations as in this section except that the biserial correlations are not homogeneous, although the averages of the biserial correlations are the same, it is noted that the results of the two tables are almost identical. This suggests that the effects of non-homogeneous biserial correlations are small, even though the non-homogeneous biserial correlations do violate the ETEM assumptions, and consequently lower the KR20 relative to the reliability.

Although the biserial correlations are not homogeneous, almost the same conclusions may be made for Tables 6.10 and 6.11 as for Tables 6.7 and 6.8 respectively;

(a) the means and standard errors of reliability estimates are almost identical in the two sets of the experiments,

(b) the F-tests are quite robust against the violation of the ANOVA model and normal distribution theory for the binary item cases if difficulty parameters are homogeneous, and

(c) the item difficulty parameters affect the distribution considerably, if they are not homogeneous, thus inflating or deflating real Type one errors for the F-tests.

TABLE 6.9

Comparisons of Calculated Test Parameters Under the Normal Ogive Model With Empirical Values Based on the Parallel Form Method, Normal Error Scores, Non-Homogeneous Biserial Correlations, NI = 30030, J = 9

Exp. No.	Tr. Dis.	Bis.	Dif.	Theoretical (N.O.) Mean (1)	Var. (2)	Rel. (3)	KR20 (4)	Observed by P.F.M. Mean (5)	Var. (6)	Rel. (7)	KR20 (8)
01	U1	B2	D1	4.5	8.160	0.820	0.815	4.520	9.131	0.853	0.848
02	NO	B2	D1	4.5	8.160	0.820	0.815	4.504	8.127	0.819	0.814
03	EX	B2	D1	4.5	8.160	0.820	0.815	4.147	7.130	0.779	0.773
04	U1	B2	D2	4.5	5.009	0.777	0.754	4.441	5.459	0.796	0.777
05	NO	B2	D2	4.5	5.009	0.777	0.754	4.507	4.984	0.776	0.754
06	EX	B2	D2	4.5	5.009	0.777	0.754	4.443	3.924	0.696	0.674
07	U1	B2	D3	6.3	6.634	0.810	0.804	6.181	7.257	0.831	0.825
08	NO	B2	D3	6.3	6.634	0.810	0.804	6.290	6.632	0.809	0.804
09	EX	B2	D3	6.3	6.634	0.810	0.804	6.211	4.723	0.672	0.667
10	U1	B2	D4	6.3	5.470	0.775	0.767	6.226	5.758	0.783	0.778
11	NO	B2	D4	6.3	5.470	0.775	0.767	6.290	5.552	0.778	0.772
12	EX	B2	D4	6.3	5.470	0.775	0.767	6.343	3.472	0.593	0.585
13	U1	B4	D1	4.5	6.474	0.739	0.734	4.503	7.109	0.775	0.769
14	NO	B4	D1	4.5	6.474	0.739	0.734	4.500	6.462	0.740	0.733
15	EX	B4	D1	4.5	6.474	0.739	0.734	4.254	5.684	0.688	0.682
16	U1	B4	D2	4.5	4.109	0.690	0.673	4.462	4.302	0.705	0.689
17	NO	B4	D2	4.5	4.109	0.690	0.673	4.495	4.099	0.691	0.674
18	EX	B4	D2	4.5	4.109	0.690	0.673	4.446	3.342	0.579	0.584
19	U1	B4	D3	6.3	5.241	0.725	0.719	6.226	5.600	0.746	0.740
20	NO	B4	D3	6.3	5.241	0.725	0.719	6.310	5.244	0.725	0.720
21	EX	B4	D3	6.3	5.241	0.725	0.719	6.221	3.854	0.570	0.564
22	U1	B4	D4	6.3	4.373	0.682	0.677	6.280	4.446	0.685	0.681
23	NO	B4	D4	6.3	4.373	0.682	0.677	6.287	4.365	0.682	0.675
24	EX	B4	D4	6.3	4.373	0.682	0.677	6.305	3.065	0.506	0.499
25	U1	B6	D1	4.5	4.072	0.510	0.503	4.475	4.218	0.533	0.525
26	NO	B6	D1	4.5	4.072	0.510	0.503	4.503	4.108	0.513	0.509
27	EX	B6	D1	4.5	4.072	0.510	0.503	4.417	3.769	0.461	0.454
28	U1	B6	D2	4.5	2.737	0.457	0.447	4.487	2.815	0.469	0.463
29	NO	B6	D2	4.5	2.737	0.457	0.447	4.500	2.731	0.454	0.444
30	EX	B6	D2	4.5	2.737	0.457	0.447	4.481	2.452	0.380	0.372
31	U1	B6	D3	6.3	3.307	0.489	0.482	6.296	3.354	0.500	0.490
32	NO	B6	D3	6.3	3.307	0.489	0.482	6.295	3.328	0.493	0.485
33	EX	B6	D3	6.3	3.307	0.489	0.482	6.275	2.782	0.361	0.357
34	U1	B6	D4	6.3	2.814	0.433	0.429	6.287	2.824	0.433	0.428
35	NO	B6	D4	6.3	2.814	0.433	0.429	6.292	2.797	0.428	0.424
36	EX	B6	D4	6.3	2.814	0.433	0.429	6.303	2.338	0.296	0.292

TABLE 6.10

Comparisons of Observed Means and Standard Errors of Reliability Estimates Under Normal Error Scores and Non-Homogeneous Biserial Correlations With the Values Obtainable From ANOVA Model and Normal Theory, N = 1000, I = 30, J = 9

Exp. No.	Tr. Dis.	Bis.	Dif.	Rel.*	Observed $\hat{\rho}$ Mean	Observed $\hat{\rho}$ S.E.	$E(\hat{\rho})$ by (5.1)-(a)	Expected S.E. by (5.3)	Expected S.E. by (5.1)-(b)
				(1)	(2)	(3)	(4)	(5)	(6)
01	U1	B2	D1	0.853	0.842	0.0436	0.842	0.0499	0.0473
02	NO	B2	D1	0.813	0.805	0.0564	0.799	0.0618	0.0600
03	EX	B2	D1	0.779	0.760	0.0739	0.763	0.0716	0.0708
04	U1	B2	D2	0.796	0.766	0.0596	0.781	0.0668	0.0653
05	NO	B2	D2	0.769	0.740	0.0707	0.752	0.0745	0.0740
06	EX	B2	D2	0.696	0.654	0.0986	0.674	0.0941	0.0975
07	U1	B2	D3	0.831	0.817	0.0505	0.819	0.0565	0.0542
08	NO	B2	D3	0.802	0.790	0.0676	0.788	0.0651	0.0635
09	EX	B2	D3	0.672	0.650	0.0988	0.647	0.1002	0.1053
10	U1	B2	D4	0.783	0.763	0.0681	0.767	0.0705	0.0695
11	NO	B2	D4	0.788	0.752	0.0859	0.772	0.0693	0.0682
12	EX	B2	D4	0.593	0.561	0.1191	0.563	0.1183	0.1306
13	U1	B4	D1	0.775	0.759	0.0630	0.759	0.0728	0.0721
14	NO	B4	D1	0.734	0.719	0.0839	0.714	0.0843	0.0855
15	EX	B4	D1	0.688	0.659	0.1088	0.665	0.0962	0.1002
16	U1	B4	D2	0.705	0.673	0.0841	0.684	0.0917	0.0945
17	NO	B4	D2	0.684	0.654	0.0998	0.660	0.0972	0.1015
18	EX	B4	D2	0.597	0.555	0.1348	0.567	0.1174	0.1293
19	U1	B4	D3	0.746	0.723	0.0838	0.727	0.0810	0.0816
20	NO	B4	D3	0.718	0.700	0.0951	0.698	0.0883	0.0904
21	EX	B4	D3	0.570	0.540	0.1278	0.539	0.1232	0.1379
22	U1	B4	D4	0.685	0.662	0.0948	0.661	0.0970	0.1013
23	NO	B4	D4	0.702	0.647	0.1185	0.680	0.0927	0.0957
24	EX	B4	D4	0.506	0.470	0.1474	0.469	0.1358	0.1586
25	U1	B6	D1	0.533	0.496	0.1448	0.498	0.1308	0.1500
26	NO	B6	D1	0.506	0.476	0.1566	0.470	0.1358	0.1585
27	EX	B6	D1	0.461	0.414	0.1791	0.421	0.1438	0.1731
28	U1	B6	D2	0.469	0.428	0.1578	0.430	0.1424	0.1704
29	NO	B6	D2	0.452	0.405	0.1729	0.411	0.1453	0.1759
30	EX	B6	D2	0.380	0.329	0.1963	0.334	0.1563	0.1991
31	U1	B6	D3	0.500	0.458	0.1585	0.463	0.1370	0.1605
32	NO	B6	D3	0.484	0.448	0.1638	0.446	0.1398	0.1656
33	EX	B6	D3	0.361	0.314	0.1962	0.313	0.1588	0.2052
34	U1	B6	D4	0.433	0.388	0.1736	0.391	0.1484	0.1821
35	NO	B6	D4	0.466	0.378	0.1901	0.427	0.1429	0.1713
36	EX	B6	D4	0.296	0.248	0.2106	0.244	0.1666	0.2259

*Theoretical values if true scores are normal, otherwise the value was obtained by the parallel form method.

(5.1) (a) $E(\hat{\rho}) = 1 - (1-\rho)\frac{I-1}{I-3}$ (b) $Var(\hat{\rho}) = (1-\rho)^2 \frac{2(I-1)(\nu+I-3)}{(J-1)(I-3)^2(I-5)}$

(5.3) $Var(\hat{\rho}) = \frac{(1-\rho^2)^2}{I}$

TABLE 6.11

Comparisons of Observed Lower and Upper 5% Critical Points Under Normal Error Scores and Non-Homogeneous Biserial Correlations With the Values Obtainable From the ANOVA Model and Normal Theory, and Real Type One Error of F-Test When Nominal Value is Fixed to the 5% Level, N = 1000, I = 30, J = 9

Exp. No.	Tr. Dis.	Bis.	Dif.	Rel.[1]	Observed C.P.[2] Lower	Upper	Theoretical C.P.[3] Lower	Upper	Real Sig. (0/0) Lower	Upper
				(1)	(2)	(3)	(4)	(5)	(6)	(7)
01	U1	B2	D1	0.853	0.765	0.902	0.754	0.903	3.40	4.60
02	NO	B2	D1	0.813	0.704	0.877	0.688	0.877	4.60	4.00
03	EX	B2	D1	0.779	0.629	0.856	0.632	0.855	5.20	5.40
04	U1	B2	D2	0.796	0.653	0.845	0.660	0.866	5.90	1.40
05	NO	B2	D2	0.769	0.598	0.828	0.615	0.848	7.10	0.80
06	EX	B2	D2	0.696	0.462	0.778	0.493	0.800	6.90	1.40
07	U1	B2	D3	0.831	0.733	0.884	0.718	0.889	3.20	3.70
08	NO	B2	D3	0.802	0.670	0.877	0.670	0.870	6.70	6.00
09	EX	B2	D3	0.672	0.465	0.780	0.452	0.784	4.20	4.20
10	U1	B2	D4	0.783	0.629	0.858	0.638	0.857	6.10	5.10
11	NO	B2	D4	0.788	0.592	0.856	0.645	0.860	7.70	5.90
12	EX	B2	D4	0.593	0.345	0.722	0.321	0.732	4.00	3.30
13	U1	B4	D1	0.775	0.648	0.845	0.625	0.852	3.10	3.40
14	NO	B4	D1	0.734	0.564	0.824	0.555	0.824	5.20	4.60
15	EX	B4	D1	0.688	0.470	0.801	0.479	0.794	5.40	6.10
16	U1	B4	D2	0.705	0.513	0.783	0.508	0.806	4.80	2.30
17	NO	B4	D2	0.684	0.482	0.777	0.472	0.791	5.20	2.10
18	EX	B4	D2	0.597	0.308	0.720	0.328	0.734	5.90	2.50
19	U1	B4	D3	0.746	0.597	0.831	0.576	0.832	5.50	4.50
20	NO	B4	D3	0.718	0.526	0.820	0.530	0.814	6.20	5.20
21	EX	B4	D3	0.570	0.304	0.717	0.283	0.717	3.80	5.00
22	U1	B4	D4	0.685	0.467	0.787	0.473	0.792	5.40	4.00
23	NO	B4	D4	0.702	0.416	0.796	0.502	0.803	7.30	6.20
24	EX	B4	D4	0.506	0.193	0.667	0.175	0.674	4.30	4.00
25	U1	B6	D1	0.533	0.220	0.682	0.220	0.692	5.10	3.90
26	NO	B6	D1	0.506	0.174	0.674	0.176	0.674	5.40	4.10
27	EX	B6	D1	0.461	0.052	0.657	0.100	0.644	6.10	5.90
28	U1	B6	D2	0.469	0.129	0.650	0.114	0.650	4.30	5.00
29	NO	B6	D2	0.452	0.085	0.635	0.085	0.639	5.20	4.50
30	EX	B6	D2	0.380	-0.028	0.594	-0.036	0.591	4.60	5.10
31	U1	B6	D3	0.500	0.165	0.657	0.165	0.670	5.00	4.30
32	NO	B6	D3	0.484	0.128	0.672	0.138	0.660	5.50	6.00
33	EX	B6	D3	0.361	-0.056	0.579	-0.068	0.578	4.40	5.10
34	U1	B6	D4	0.433	0.063	0.629	0.053	0.626	4.40	5.30
35	NO	B6	D4	0.466	0.026	0.624	0.109	0.648	5.80	5.20
36	EX	B6	D4	0.296	-0.146	0.525	-0.175	0.536	4.50	3.50

[1]Theoretical values if true scores are normal, otherwise the values obtained by the parallel form method.

[2]Observed lower and upper 5% critical points of the distribution β.

[3]Theoretical lower and upper 5% critical points of the distribution of β under ANOVA and normal theory.

From the above observations, the following conclusions are tentatively made.

(a) The non-homogeneous biserial correlation distorts the distribution slightly to the left for high reliability cases, but the differences are not substantial.

(b) The effects of non-homogeneous biserial correlations on expected values and standard errors of reliability estimates are minimal, and formulas (5.1) and (5.3) are quite satisfactory.

(c) The effects of non-homogeneous biserial correlations on test parameters are minimal.

(d) The F-tests are robust for binary item test cases if the difficulty parameters are homogeneous.

6.5 Summary

In order to investigate the effects of non-normal latent and error scores, non-homogeneous difficulty parameters and biserial correlations on the sampling distribution of reliability estimates based on formula (2.13), altogether 96 experiments were performed by REL02 using various combinations of distribution parameter sets with $N = 1000$, $I = 30$, and $J = 9$. The findings in this chapter may be summarized as the following:

(a) The effects of non-normal distribution of error scores $\{\varepsilon_{ij}\}$ in terms of response strength variables $\{y_{ij}\}$ are negligible, as was the case for the continuous part score cases in Chapter Five.

(b) The non-normal latent scores affect the population parameters such as variance, reliability and KR20. The normal ogive model underestimates these parameters for the uniform latent score distribution, and overestimates them for the exponential case.

(c) Formulas (5.1) and (5.3) are quite satisfactory for binary item cases; formula (5.1)-(b) seems superior to (5.3) for the calculation of the standard error of reliability estimates.

(d) The item difficulty parameters are the most important factor for the distribution of reliability estimates. They will affect the test score variance, reliability and KR20. The non-homogeneous difficulty sets give lower values for these parameters.

(e) The item difficulty parameters systematically affect the distribution of reliability estimates. The non-homogeneous difficulty sets shift the distribution leftward.

(f) The effect of non-homogeneous biserial correlations are negligible if the heterogeneity is moderate.

(g) The F-test based on (2.17) is robust if any one of the following conditions is satisfied.

i) Relibability is low, i.e., ρ is close to zero.

ii) Only lower portions of the sampling distribution of reliability estimates are used for the inference, namely the null hypothesis is directional, being bounded only by the lower end.

iii) The difficulty parameters are almost homogenous.

(h) The difficulty parameter sets may deflate the real Type one errors if they are not homogeneous for inference which uses only upper tail of the sampling distribution.

CHAPTER SEVEN

SUMMARY, IMPLICATIONS, EXAMPLES OF APPLICATION, AND RECOMMENDATIONS

7.1.0 Summary of Findings

The purpose of this study was (a) to review the more liberal concepts of test reliability theory in terms of models and assumptions underlying them, (b) to examine the sampling distribution of reliability estimates based on Alpha or KR20 formulas using these models with various combinations of the distribution of true and error scores, and (c) to compare the empirical distributions thus obtained by computer simulation under these model-distribution combinations with those obtainable theoretically under a mixed model ANOVA and normal theory Using computer simulated hypothetical test score matrices, a number of statistical sampling experiments were performed to obtain empirical distributions, and some analytical means were also employed to obtain a new formula for the standard error of reliability estimates. Findings in this study are summarized in the following three sections.

7.1.1 Test Models

(a) The most general model for the continuous part test score is found to be the multi-factor true score model. The model includes other more restrictive models as special cases. By imposing a uni-factor true score constraint, the model becomes a congeneric true score model. If homogeneity of true score variance is assumed, the congeneric model becomes essentially τ equivalent measurement. The latter model includes the ANOVA or essentially parallel measurement model as a special

case with the additional assumption of the homogeneity of error variances. The classical parallel test model is a special case of ANOVA model, namely the means of part test scores are all equal. The Alpha coefficient is equal to the reliability if, and only if the essentially τ equivalent measurements condition is satisfied, otherwise it is in general lower than the reliability. The sampling distribution of reliability estimates is known only for the case of the ANOVA model and normality assumptions of true and error scores.

(b) For the binary item test case, a similar model as the continuous case may be considered for the hypothetical 'response strength' variable. A mathematical model and distributional assumptions are required to associate the response strength variable to the observed item scores. Under the normal ogive model, the test parameters such as variance, reliability, and KR20 are amenable for calculation by means of numerical methods if the item parameters, such as biserial correlation and difficulty parameters, are specified. The essentially τ equivalent measurement assumption is satisfied if and only if all biserial correlations and difficulty parameters are equal, i.e., all items are homogeneous; otherwise KR20 is lower than reliability. The sampling distribution of reliability estimates for binary item test is not yet known, except by approximation using the ANOVA model and normal theory.

7.1.2 Sampling Distribution Under Various Models and Assumptions

(a) Applying Tukey's result, a new formula for the standard error of reliability estimate was derived. The formula depends only on sample size, number of part tests, reliability, and the kurtosis

of the test scores, and is found to be superior to the traditional formula based on normal theory when the distribution of true score is not normal.

(b) The effects of non-normal error scores distributions are found to be negligible for not too small J, the number of part tests or items, for both the continuous and binary cases.

(c) For continuous test score cases, the effects of non-normal distributions of true scores are found to be significant, i.e., the distribution of reliability is systematically distorted. If the essentially τ equivalent assumption is not satisfied, the distribution is systematically shifted leftward or to the lower direction of reliability. This effect is more clearly observed for the multi-factor true score model case indicating inappropriateness of the Alpha formula for the model. The effects of non-homogeneous error variance were found to be negligible.

(d) For the binary item case the effect of non-normal distributions of latent scores is not so obvious. The formula for the standard error derived under the ANOVA model and normal theory seems quite robust against violation of assumptions imposed by a binary scoring scheme. The test parameters depend on the shape of latent score distributions for fixed biserial correlation and difficulty parameters. If the essentially τ equivalent measurement assumption is not satisfied, i.e., biserial correlation and/or difficulty parameters are not all homogeneous, the distribution of reliability estimates is shifted leftward systematically. The effects of non-homogeneous difficulty parameters seems more severe than that of non-homogeneous biserial correlations.

7.1.3 Robustness of F-Test

The F-test based on ANOVA model and normal theory is robust against violation of the following assumptions:

(a) Normality of error scores for both continuous and binary cases.

(b) Homogeneity of error score variances for continuous cases.

(c) Homogeneity of biserial correlations for the binary case if the violation is not too extreme.

(d) Normality of latent score distributions for binary case.

The F-test may be misleading if the following conditions are not satisfied.

(a) Uni-factorness of true and latent score distributions.

(b) Normality of true scores for continuous case. Especially positive kurtosis of true scores results in severe distortions.

(c) Essentially τ equivalent assumptions (approximately at least).

(d) Homogeneity of item difficulty parameters (approximately at least).

Nevertheless, in all cases, the F-test is robust against any violation of assumptions if the population reliability is close to zero for both continuous and binary cases. If only the lower tail portion of the distribution is used for the binary item test, the significance test is also robust in most cases.

7.2.0 Implications to Test Theory and Applications

In this study, it has been demonstrated that the distribution of the reliability estimate depends significantly on the models employed,

the underlying assumptions, and the parameters of part tests or items. Therefore the validity of any statistical inference about reliability largely depends on the validity of models and assumptions like any other statistical inference. Therefore it is essential, for statistical inference about reliability, to know the models appropriate for the test in use, and the population characteristics for the test must be known a priori. For a casual user of psychological and educational tests, this is an almost impossible task. Therefore, for test users and/or other researchers, the findings in this study may not be of any practical use without knowledge of the above information about the test except when robust conditions are present.

However, for a test author, or for a test reviewer, the task of gathering the necessary data may be accomplished as a by-product of the usual procedure for the test development, since an administration of the test to a comparatively large sample of subjects from the population for whom the test is developed is usually involved in order to standardize and to obtain test norms. The test statistics based on such large samples may be used to obtain such information.

Although there is no agreed upon statistical and psychometric methods to obtain such parameters, some efficient methods for the calculation of part test parameters have been developed recently by a number of psychometricians.

For example, Kristof (1969) considered the estimation of the true score variance σ_A^2 and error score variance $\{\sigma_{ej}^2\}$ under an essentially τ equivalent measurement assumption by employing maximum likelihood method. He derived the likelihood equations and found that these could be solved rapidly by a simple Newton-Raphson procedure.

For the binary item test cases, the item difficulty is easily

calculated, and the biserial correlation parameters may be obtained by factor analysis of the tetrachoric correlation matrix from the results of (3.14), if the latent score has a uni-factor structure.

Jöreskog (1971) has shown some examples of model identification techniques by employing maximum likelihood factor analysis on the disperson matrices of test scores obtained from large samples.

In regard to distributions, the distribution of error score is found to be not important, but the shape of the distribution of true scores can affect the reliability estimate significantly. Although the distribution of true scores is not observable directly, since only the kurtosis of true score will affect the distribution of reliability estimates, and it can be indirectly evaluated by the test score kurtosis divided by the square of reliability from the results of (5.9), the normality of true score may be investigated partly by examining the test score kurtosis if it is obtained from a large sample.

Therefore, a test author or reviewer would be doing a service to the users of a test, if he provided information about the model involved, and distributions and parameter values in the population for which the test is developed. If the test satisfies the ANOVA model and normal theory assumptions, or violates only those assumptions which are known to be unimportant, the author or reviewer may recommend the use of the F-test for the inference about reliability. In this case the author or reviewer needs to supply only the information about the population reliability. Otherwise, the author or reviewer should either provide all information necessary for simulation of such tests by the computer program developed in this study, or alternatively, provide a table of upper and lower critical points of the distribution of

reliability estimates as a function of sample size, and should probably also provide the values of the standard errors. Then the user could easily determine whether the observed reliability is significantly different or not from the population value at a specific significance level.

7.3.0 Example 1: Application to Continuous Case

Since it was not possible to find an appropriate example of a test and its manuals which provide the necessary information for the test models and the other information necessary for the application of computer simulation techniques, somewhat arbitrary example data were selected to show how the findings in this study and the computer programs developed might be applied in a practical situation.

Jöreskog (1971) analyzed a dispersion matrix based on four measures used by Votaw (1948) to establish methods of obtaining reader reliability in essay scoring for an English composition test, and identified the model as a congeneric true score model. The dispersion matrix was obtained from 126 subjects, and is given in Table 7.1.

TABLE 7.1

Dispersion Matrix of Votaw's Essay Test Data,
I = 126

Measure	1	2	3	4
1	25.0704	12.4363	11.7257	20.7510
2	12.4363	28.2021	9.2281	11.9732
3	11.7257	9.2281	22.7390	12.0692
4	20.7510	11.9732	12.0692	21.8707

He employed maximum likelihood factor analysis, and gave the estimate of the standard deviation of true score or factor loading as,

$$\underline{\lambda}' = [4.57 \quad 2.68 \quad 2.65 \quad 4.53] \ .$$

Therefore, if a test author published a test consisting of four part tests and obtained the same results as above based on a large sample, these values may be regarded as population parameter values if small discrepancies in covariance terms are ignored. Then the test score model would be as follows,

$$\underline{y}_i = \begin{bmatrix} 4.57 \\ 2.68 \\ 2.65 \\ 4.53 \end{bmatrix} \begin{bmatrix} f_i \end{bmatrix} + \begin{bmatrix} 2.0459 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 4.5847 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 3.9644 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 1.1618 \end{bmatrix} \begin{bmatrix} \underline{\varepsilon}_i \end{bmatrix} \ .$$

Then,

$$D(\underline{y}_i) = \underline{\Sigma} = \underline{\lambda}\ \underline{\lambda}' + \underline{\Psi}^2 = \begin{bmatrix} 25.0704 & 12.2476 & 12.1105 & 20.7021 \\ 12.2476 & 28.2021 & 7.1020 & 12.1404 \\ 12.1105 & 7.1020 & 22.7390 & 12.0045 \\ 20.7021 & 12.1404 & 12.0045 & 21.8707 \end{bmatrix},$$

and,

$$\text{Alpha} = 0.812329, \qquad \rho = 0.831249.$$

Since the assumption of the homogeneity of true score variances is violated, the essentially τ equivalent measurement assumption is not valid and hence the Alpha coefficient is lower than reliability as expected.

From the findings of this study, it is known that the effects of non-homogeneous true score variance is not too great with moderate differences among the elements of the factor loading vector $\underline{\lambda}$, but

the differences for this data seem exceptionally large and also the difference between Alpha coefficient and the reliability is substantial. Therefore a systematic distortion of the distribution of reliability estimates toward lower reliability is expected. Seven computer simulation experiments were performed with the Jöreskog's model with $N = 2000$ and assumed normality of true and error scores. Both estimation formulas, namely the Alpha formula of (2.13) and Kristof's unbiased formula of (5.2)-(a) were used for estimation of sample reliability. Observed upper and lower 5% critical points together with standard errors are summarized in Table 7.2. The observed values are also compared with those obtainable under the ANOVA model and normal theory. The sample sizes I, the number of subjects, used for these experiments are 10, 15, 20, 25, 30, 35, and 40 respectively. Figures 7.1 - 7.7 compares empirical distribution with the theoretical distributions indicating the effect of the violation of the essentially τ equivalent measurement assumptions.

A table similar to Table 7.2 might accompany the test manuals or test review report so that test users may consult the table whenever they make inferences about the reliability. For example, if a teacher administered the test to a sample of 20 students and obtained $\hat{\rho} = 0.892$, then by consulting this table she may conclude that the difference between the population value 0.812 and her sample value is not significant at 5% level of Type one error. Therefore, she may not claim that her sample is significantly different from the population for which the test is developed as far as the reliability is concerned. The author or researcher could develop a slightly different table if the population test score is not normal. For

TABLE 7.2

Lower and Upper 5% Critical Points of the Distribution of Reliability Estimates Votaw-Jöreskog Data, Normal True and Error Score Distributions, Congeneric Model, $\rho = 0.8313$, Alpha = 0.8123, N = 2000

Est. Formula	I	Observed Lower 5%	Observed Upper 5%	Observed S.E.	Theoretical Under ANOVA Lower 5%	Theoretical Under ANOVA Upper 5%	Theoretical Under ANOVA S.E.
Alpha (2.13) Biased	10	0.4501	0.9223	0.1671	0.7141	0.9250	0.1540
	15	0.5565	0.9048	0.1130	0.6186	0.9128	0.0998
	20	0.6194	0.8952	0.0881	0.6651	0.9048	0.0785
	25	0.6586	0.8917	0.0733	0.6920	0.8989	0.0665
	30	0.6872	0.8841	0.0624	0.7098	0.8944	0.0587
	35	0.6918	0.8808	0.0608	0.7226	0.8907	0.0531
	40	0.6952	0.8763	0.0571	0.7323	0.8877	0.0488
Kristof's (5.2)-(a)	10	0.5723	0.9395	0.1301	0.6221	0.9417	0.1198
	15	0.6199	0.9184	0.0970	0.6731	0.9252	0.0856
	20	0.6595	0.9062	0.0790	0.7003	0.9148	0.0702
	25	0.6871	0.9007	0.0673	0.7176	0.9073	0.0610
	30	0.7088	0.8921	0.0582	0.7298	0.9017	0.0546
	35	0.7099	0.8878	0.0572	0.7389	0.8972	0.0499
	40	0.7109	0.8826	0.0542	0.7460	0.8934	0.0463

I : sample size, i.e., number of subjects.

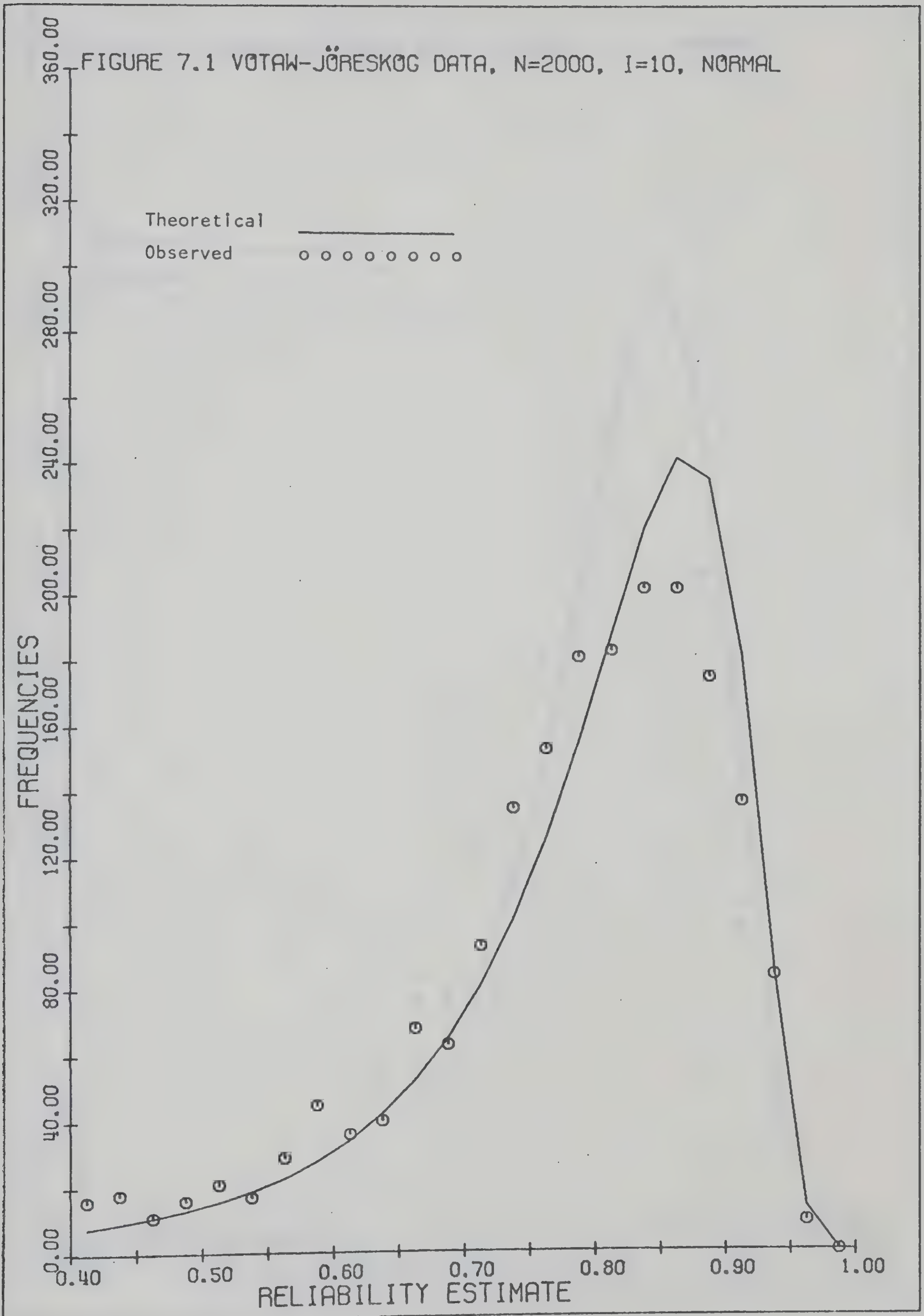


FIGURE 7.1 VOTAW-JÖRESKOG DATA, N=2000, I=10, NORMAL
Theoretical
Observed
FREQUENCIES
0.00
40.00
80.00
120.00
160.00
200.00
240.00
280.00
320.00
360.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

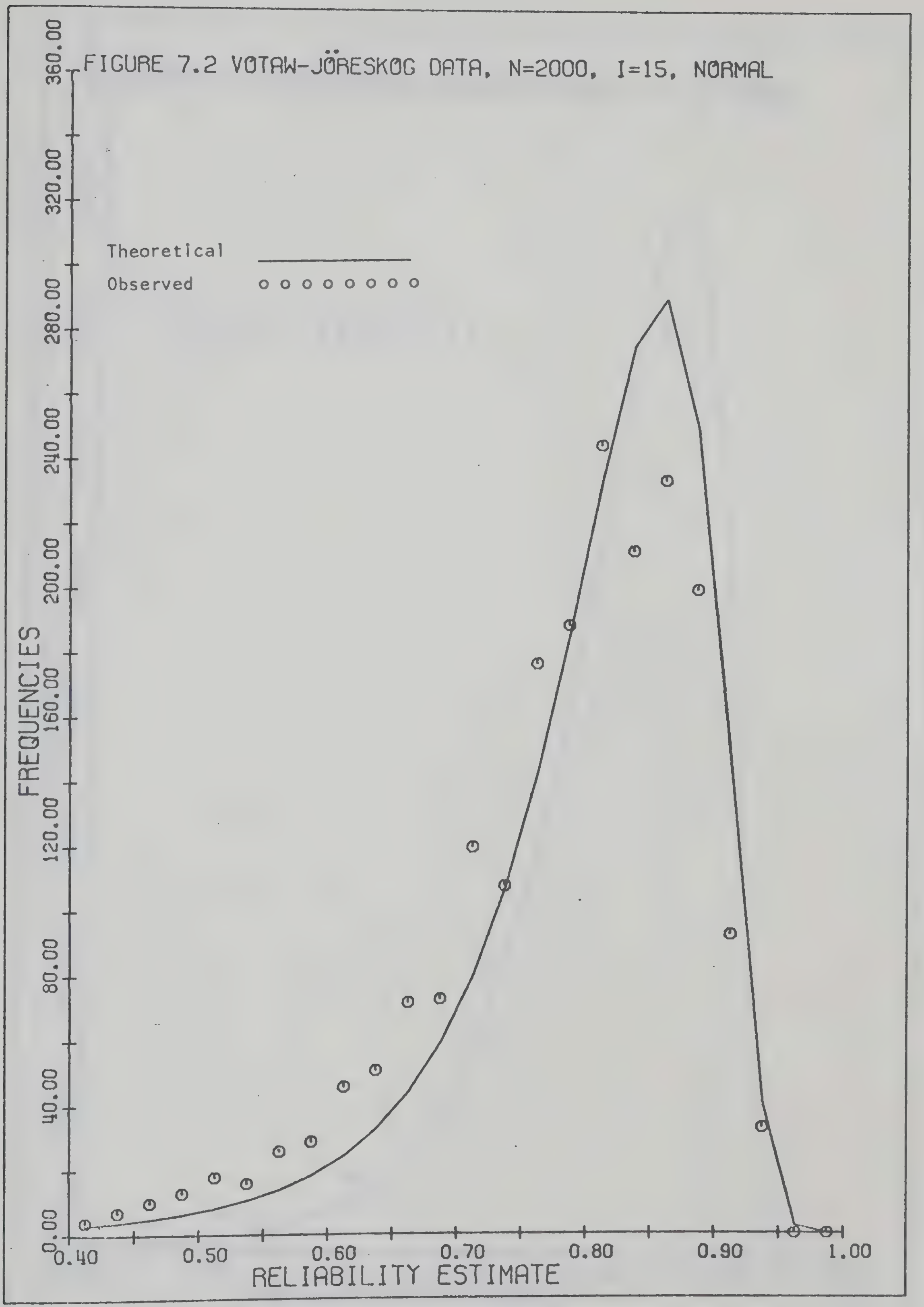
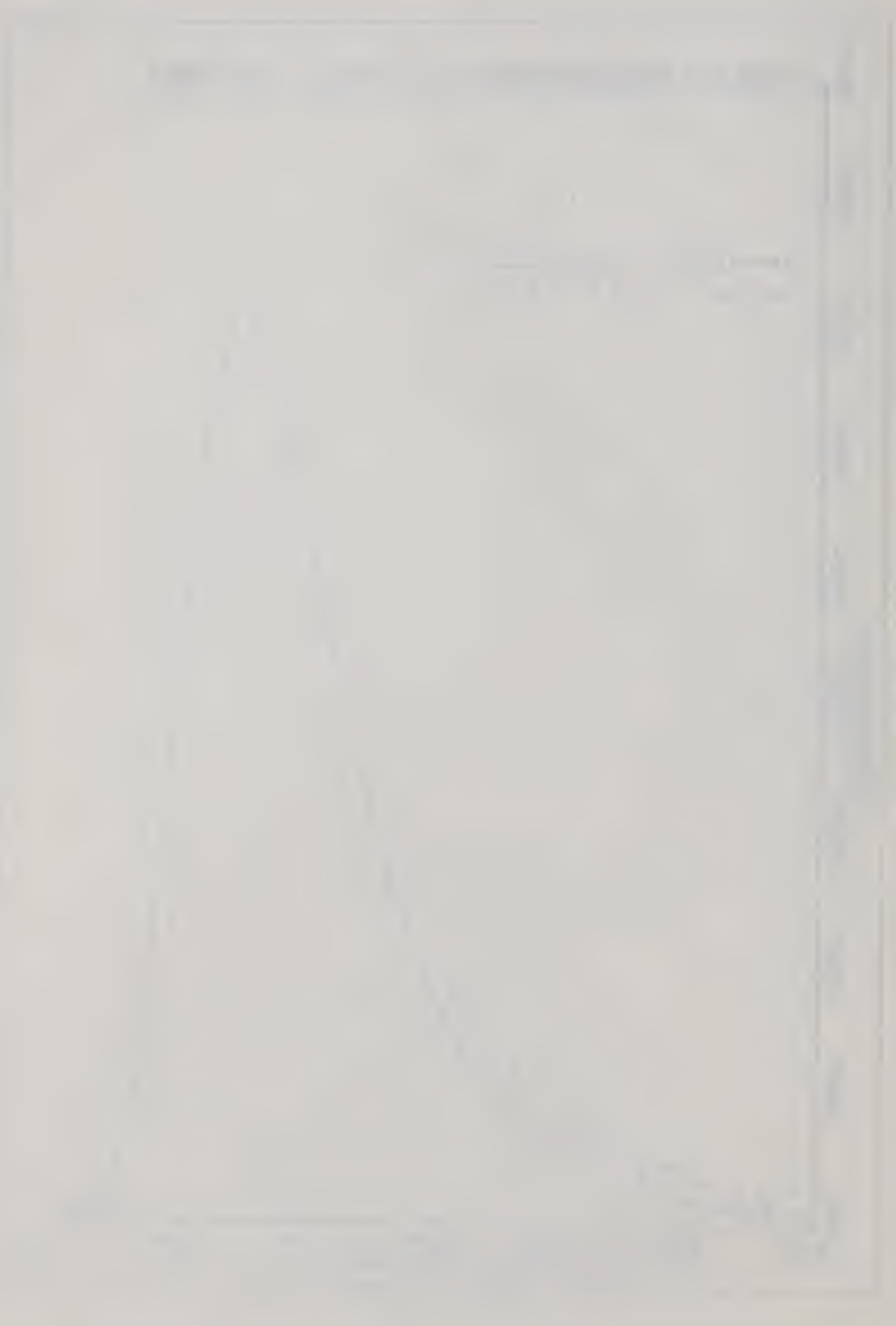


FIGURE 7.2 VOTAW-JÖRESKOG DATA, N=2000, I=15, NORMAL
Theoretical
Observed
FREQUENCIES
RELIABILITY ESTIMATE
0.00
40.00
80.00
120.00
160.00
200.00
240.00
280.00
320.00
360.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00



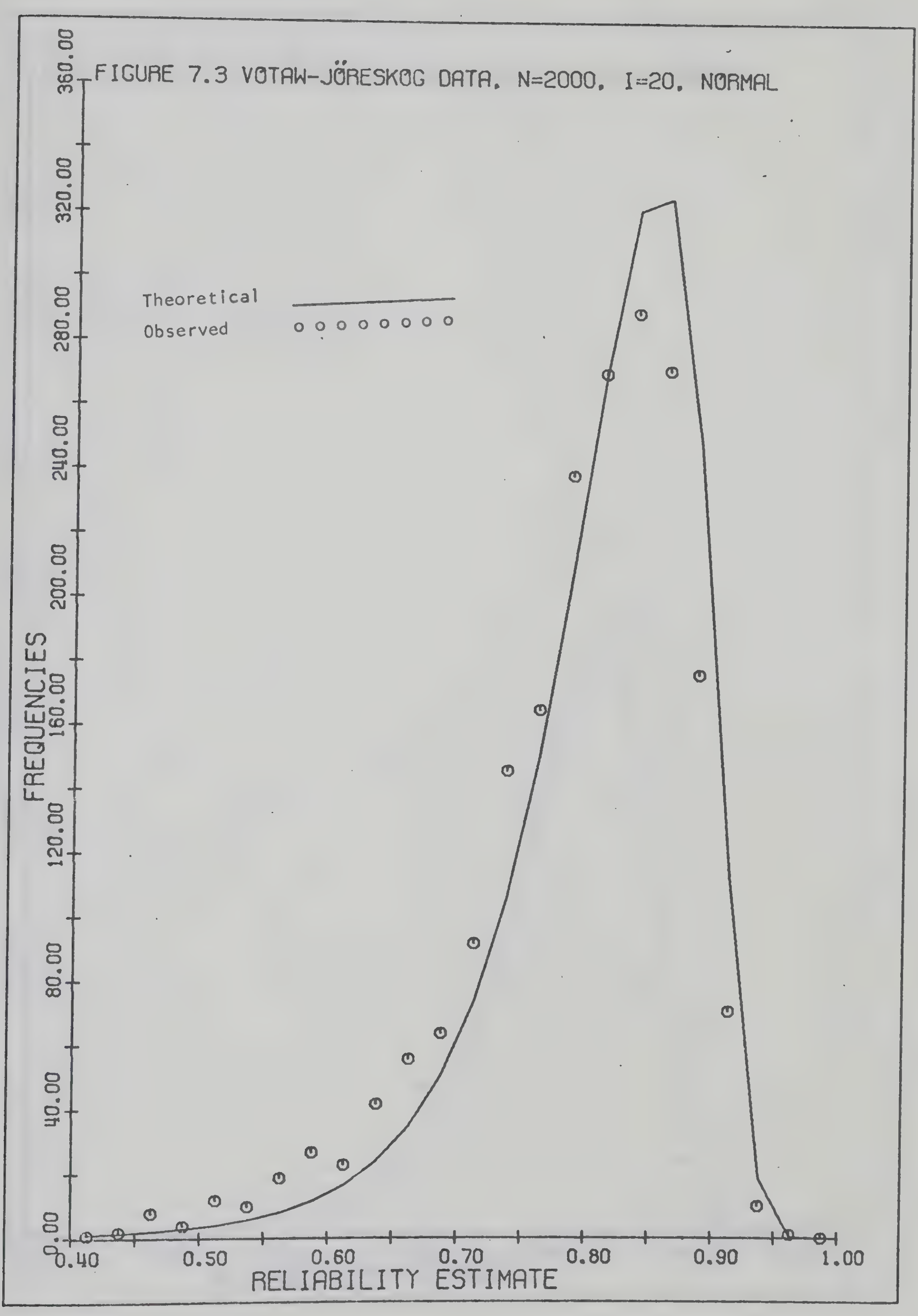


FIGURE 7.3 VOTAW-JÖRESKOG DATA, N=2000, I=20, NORMAL
Theoretical
Observed
FREQUENCIES
0.00
40.00
80.00
120.00
160.00
200.00
240.00
280.00
320.00
360.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

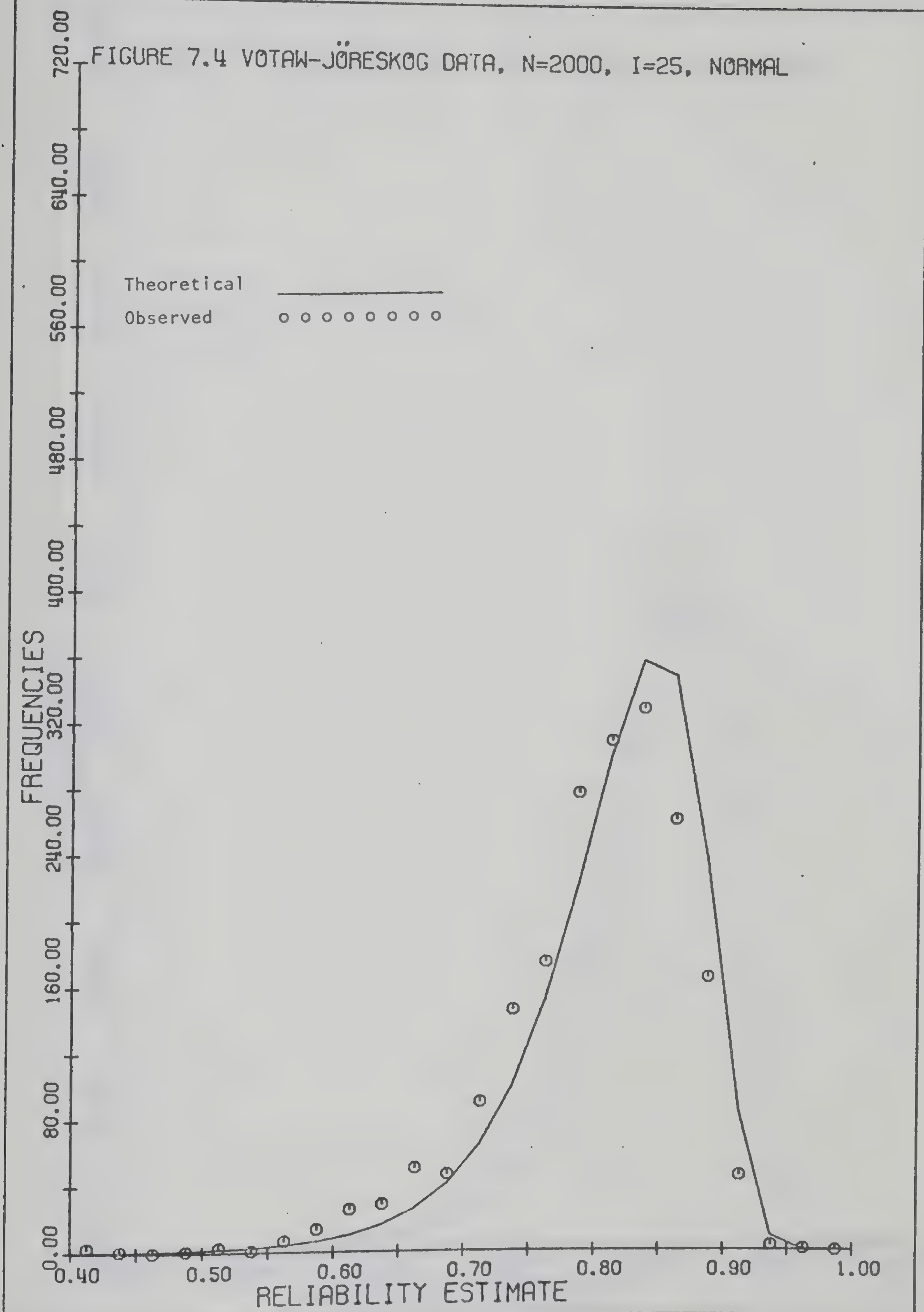


FIGURE 7.4 VOTAW-JÖRESKOG DATA, N=2000, I=25, NORMAL
Theoretical
Observed
FREQUENCIES
0.00
80.00
160.00
240.00
320.00
400.00
480.00
560.00
640.00
720.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

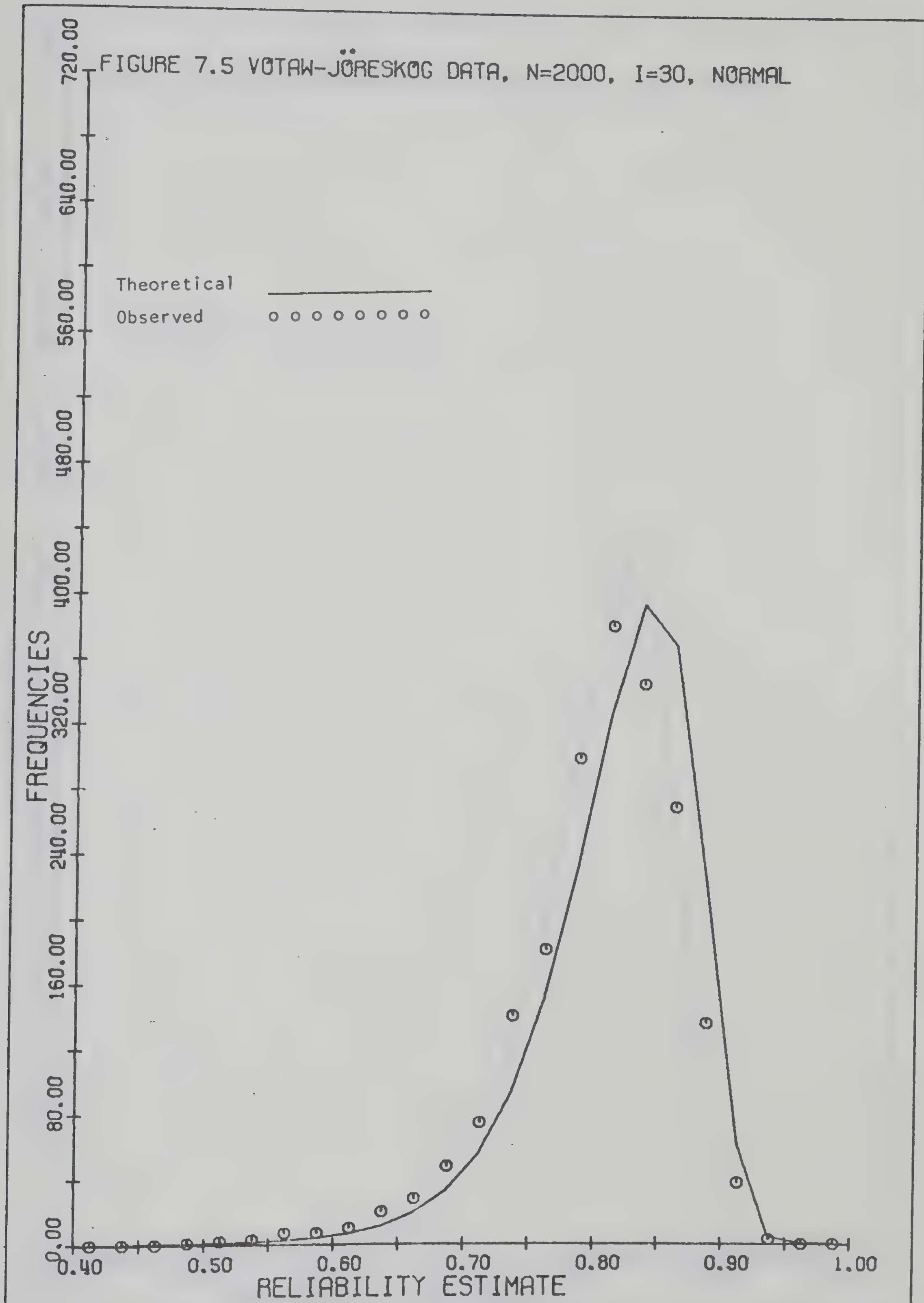


FIGURE 7.5 VOTAW-JÖRESKOG DATA, N=2000, I=30, NORMAL
Theoretical
Observed
FREQUENCIES
0.00
80.00
160.00
240.00
320.00
400.00
480.00
560.00
640.00
720.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

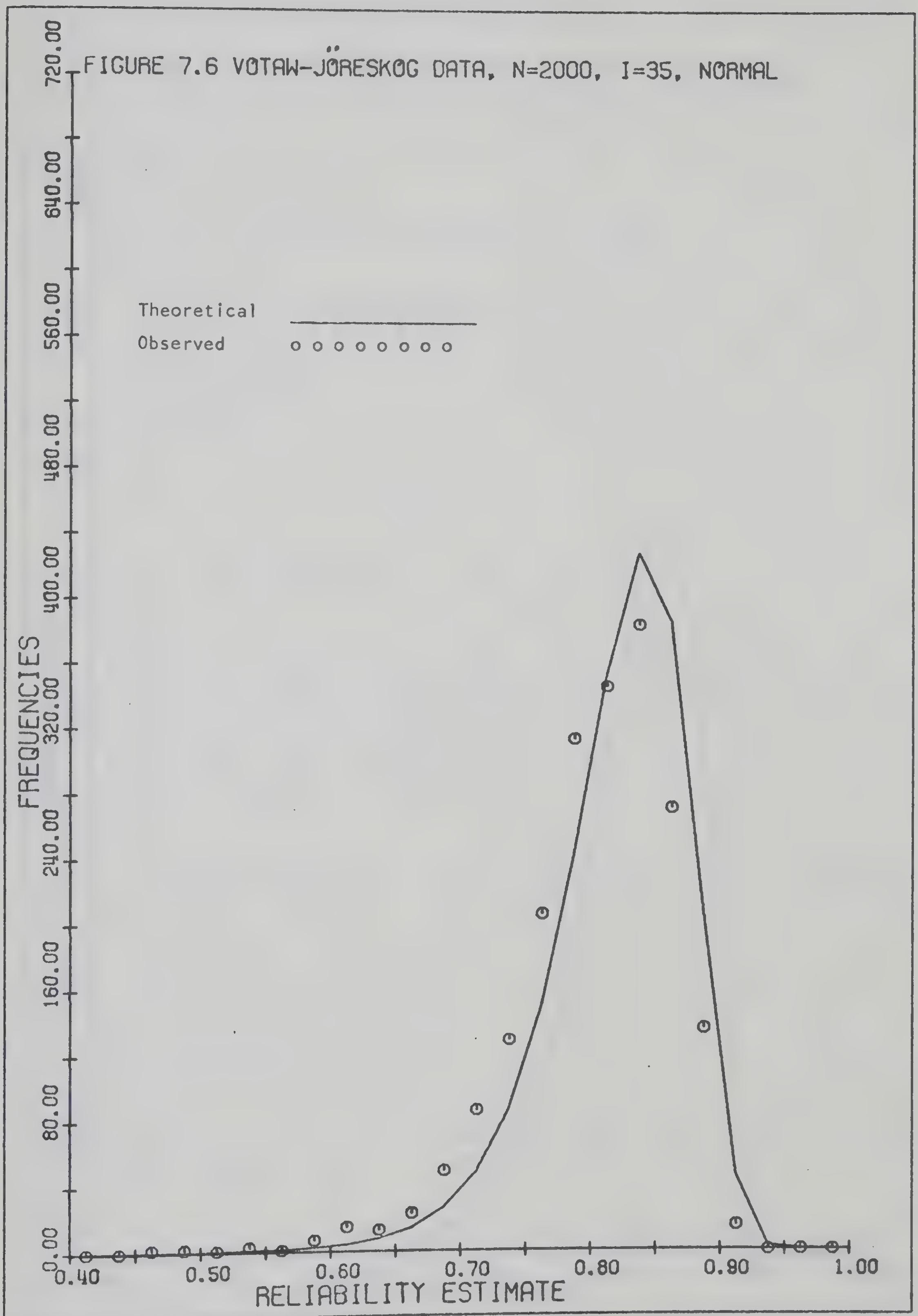


FIGURE 7.6 VOTAW-JÖRESKOG DATA, N=2000, I=35, NORMAL
Theoretical
Observed
FREQUENCIES
0.00
80.00
160.00
240.00
320.00
400.00
480.00
560.00
640.00
720.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

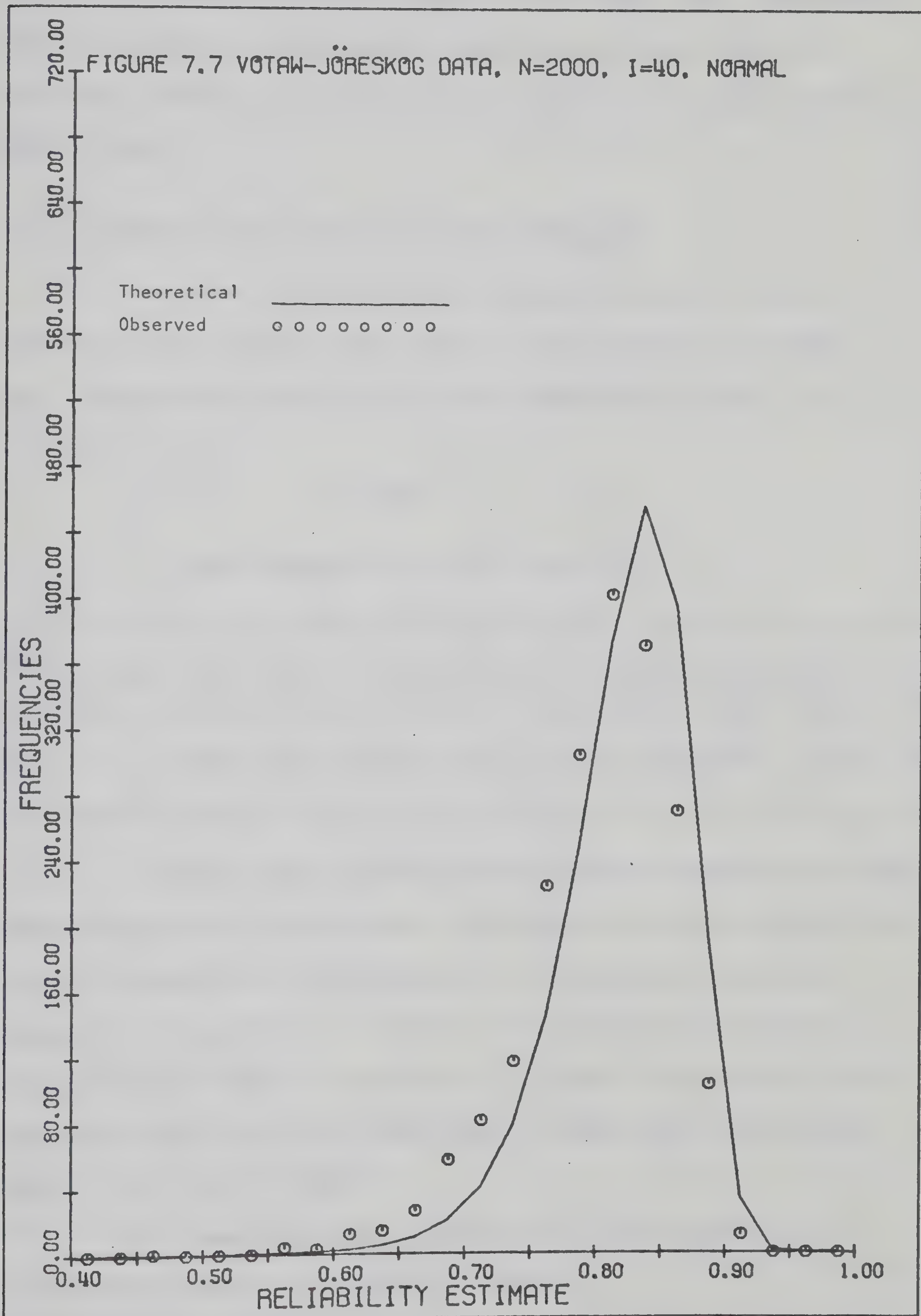


FIGURE 7.7 VOTAW-JÖRESKOG DATA, N=2000, I=40, NORMAL
Theoretical
Observed
FREQUENCIES
0.00
80.00
160.00
240.00
320.00
400.00
480.00
560.00
640.00
720.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

example, if ranked marks were assigned for each part test, the greater likelihood is that the true scores would be distributed uniformly rather than normally, and the shapes of reliability estimates would be much different.

7.4.0 Example 2: Application to Binary Item Case

A hypothetical binary item test consisting of 9 items is considered as an example. The values of item parameters are taken from Lord and Novick (1968, p. 379), and summarized in Table 7.3.

TABLE 7.3

Item Parameters of a Nine Item Test

Items	1	2	3	4	5	6	7	8	9
Difficulty	0.096	0.199	0.338	0.434	0.471	0.574	0.676	0.801	0.822
Biserial Cor.	0.490	0.717	0.549	0.593	0.595	0.640	0.476	0.530	0.495

It may be noted that the item difficulty parameters are rather heterogeneous with a value as small as 0.096 to as high as 0.822. Therefor it is expected that the essentially τ equivalent measurement assumption is substantially violated. To investigate the sampling distribution of reliability estimates of a binary item test with these parameters under the normal ogive model, an experiment was performed with $I = 30$, and $N = 1000$.

The theoretical test parameters and those obtained by parallel form method are compared in Table 7.4.

TABLE 7.4

Test Parameters of a Nine Item Test

Methods	Mean	Variance	Reliability	KR20
Theoretical	4.4710	4.0054	0.6632	0.6498
Parallel Form	4.4774	4.0023	0.6654	0.6526

Therefore a user of this test may compare her observed test mean, variance, and KR20 with the values given in this table, and can make some conclusions about her sample group.

The shape of the distribution of reliability estimate based on (2.13) is compared with the theoretical distribution under the ANOVA and normal theory model in Figure 7.8. The distribution shows a systematic shift leftward probably due to heterogeneous difficulty parameters. The lower and upper 5% critical points of this distribution are 0.4412 and 0.7793 respectively while the theoretical values are 0.4466 and 0.7656 respectively. Therefore, if a user of the test found a reliability estimate of 0.79 with $I = 30$, it may be concluded that the reliability is significantly higher than the population value at the 5% level of significance.

7.5.0 Recommendations

As noted in Section 4.8 of Chapter Four, in the discussion of the methodological limitation of this study, the computer simulation experiments cannot be exhaustive and cover all possible combinations of models, parameters, and distributional assumptions. Also due to the limitations imposed by limited funds available for the computing

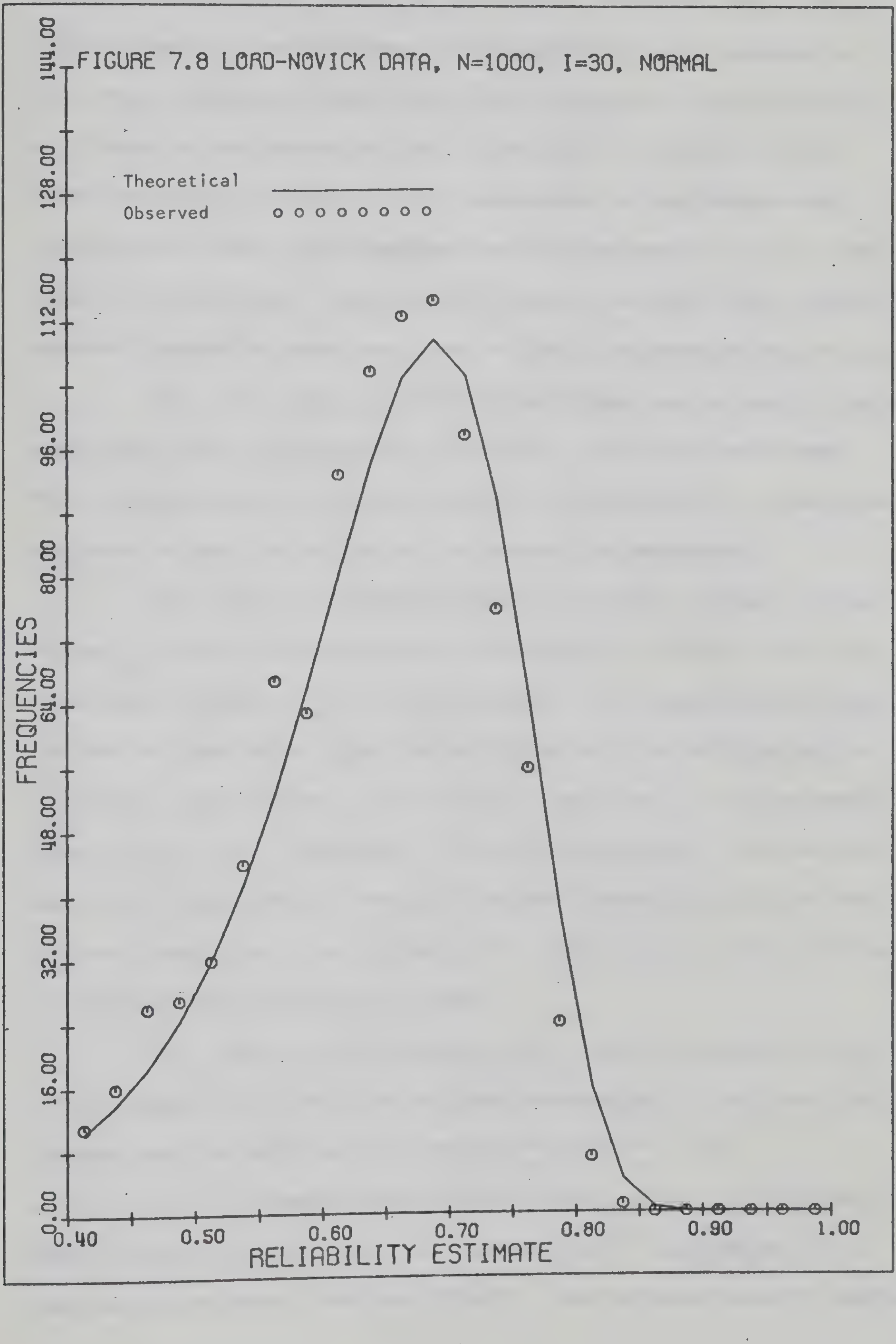


FIGURE 7.8 LORD-NOVICK DATA, N=1000, I=30, NORMAL
Theoretical
Observed
FREQUENCIES
0.00
16.00
32.00
48.00
64.00
80.00
96.00
112.00
128.00
144.00
0.40
0.50
0.60
0.70
0.80
0.90
1.00
RELIABILITY ESTIMATE

charges, the scope and extent of experiments have been restricted to certain special cases which may not always be directly relevant to real data. Because of these facts, the findings of this study will be limited to some extent in their generalization and application. Therefore, the findings will, by circumstance, be exploratory and illustrative rather than comprehensive with the emphasis having been placed on methodology. Based on the findings and experience with the computer simulation techniques, the following recommendations are made:

(a) The computer simulation techniques can be used to solve many statistical and psychometric problems in test and measurement theory and application. Further use of this technique is recommended and research must be carried out to improve the methodology.

(b) Authors of published tests, or their reviewers, should attempt to specify the appropriate test model for a given test, and place such information in the test manuals. The manuals should also include the population dispersion or tetrachoric correlation matrix of true or latent scores or estimate of them as well as the parameter values such as error variances, difficulty and biserial correlations based on a large sample. The distributional characteristics of true or latent scores and error scores of the population for which the test is developed should also be included.

(c) Some of the findings in this study are based on only a few parameter sets and distributional assumptions. Therefore, the findings must be confirmed by replicated studies with a wider range of parameter sets and with distributions of different shapes of true or latent and error scores, and if applicable, using real test score data. More specifically, the following aspects require

further investigation:

(i) The effect of non-homogeneous true score variance, i.e., the distribution of reliability estimates under the congeneric model for a wider range of $\underline{\lambda}$'s.

(ii) The effects of non-homogeneous item difficulty parameters for the binary item test cases.

(d) In this study, the size of sample is artificially fixed at $I = 30$. An investigation must be made to examine the effects of sample size to see how fast the estimate converges to its expected value with increasing sample size.

(e) The investigation of this study was limited to Type 1 sampling situations only, but a similar method can be employed for Type 2 or Type 12 sampling situations possibly with a different type of ANOVA model.

(f) The test score used in this study was a simple unweighted sum of J part test or item scores, although a weighted sum could have been easily employed. The effects of a weighted sum on the reliability estimate must be explored as an extension of this study.

(g) In this study, one of the basic assumptions of test theory was assumed to be always valid. The assumption was one of independence of error and true scores. In practice this may be violated and the effects of such violation on the distribution of the estimate of reliability must be investigated. Computer simulation would provide an ideal method for such an investigation.

(h) It is clear that Alpha coefficient as an estimate of reliability is inappropriate if the essentially τ equivalent

measurement assumption is violated too much. Therefore a new effort is necessary to find an appropriate means to estimate reliability under this condition, especially for the case of the multi-factor true score model.

(g) In this study, the investigations were limited to one sample and one reliability estimate cases, and comparisons of the estimate to the population value, but similar methods may be applied to investigate for the cases of more than one sample or reliability estimates either based on independent samples or repeated measures on the same sample to investigate the sampling distribution of the differences of the reliability estimates.

REFERENCES

ANDERSON, T.W., An Introduction to Multivariate Statistical Analysis, New York: John Wiley and Sons, 1962.

AOYAMA, Hirojiro, 'Sampling Fluctuations of the Test Reliability', Tokei-Suri Kenyojo, Tokyo Annals, Vol. 8, 1957, pp. 129-143.

BAKER, F.B., Empirical Determination of Sampling Distribution of Item Discrimination Indices and a Reliability Coefficient, Madison: University of Wisconsin, 1962.

BAY, K., Application of Linear Model Simulator to Educational and Psychological Research, A Technical Note on a Computer Program, Unpublished paper, Division of Educational Research Services, The University of Alberta, Edmonton, 1970.

BIRNBAUM, A., Statistical Theory for Logistic Mental Test Models with a Priori Distribution of Ability, Research Bulletin RB 67-12, Educational Testing Service, 1967.

BIRNBAUM, A., 'Some Latent Trait Models and Their Use in Inferring an Examinee's Ability', Cited by F.M. Lord and M.R. Novick, Statistical Theories of Mental Test Scores, Massachusetts: Addison-Wesley, 1968, Part 5.

BOCK, R.D., and M. Lieberman, 'Fitting a Response Model for n Dichotomously Scored Items', Psychometrika, Vol. 35, pp. 179-197, 1970.

BONEAU, C.A., 'The Effects of Violation of Assumptions Underlying the t-test', Psychological Bulletin, Vol. 57, pp. 49-64, 1960.

BOX, G.E.P., and M.E. Muller, 'A Note on the Generation of Random Normal Deviates', Annals of Mathematical Statistics, Vol. 29, pp. 610-611.

BROWNE, M.W., 'Fitting the Factor Analysis Model', Psychometrika, Vol. 34, 1969, pp. 375-405.

BURT, Cyril, 'Test Reliability Estimated by Analysis of Variance', The British Journal of Statistical Psychology, Vol. 8, pp. 103-118, 1955.

CHURCHMAN, C.W., 'An Analysis of the Concept of Simulation', Symposium on Simulation Models, Edited by Austin C. Hoggatt and F.E. Balderson. Cincinnati: South-Western Co., p. 12, 1963.

CLEARY, T.A., and R.L. Linn, Variability of Kuder-Richardson Formula 20 Reliability Estimates, Research Bulletin RB 68-7, Educational Testing Service, 1968.

CRONBACH, L.J., 'Coefficient Alpha and the Internal Structure of Tests', Psychometrika, Vol. 16, pp. 297-334, 1951.

CRONBACH, L.J., H. Ikeda, and R.A. Avner, 'Intra-class Correlation as an Approximation to the Coefficient of Generalizability', Psychological Reports, Vol. 15, pp. 727-736, 1964.

CRONBACH, L.J., and P.E. Meehl, 'Construct Validity in Psychological Tests', Psychological Bulletin, Vol. 52, pp. 281-301, 1955.

CRONBACH, L.J., N. Rajaratnam, and G.C. Gleser, 'Theory of Generalizability: a Liberalization of Reliability Theory', The British Journal of Statistical Psychology, Vol. 16, pp. 137-163, 1963.

CURETON, E.E., 'The Definition and Estimation of Test Reliability', Educational and Psychological Measurement, Vol. 18, pp. 715-738, 1958.

EBEL, R.L., 'Estimation of the Reliability of Ratings', Psychometrika, Vol. 16, pp. 407-424, 1951.

FELDT, L.S., 'The Approximate Sampling Distribution of Kuder-Richardson Reliability Coefficient Twenty', Psychometrika, Vol. 30, pp. 357-370, 1965.

FELDT, L.S., 'A Test of the Hypothesis that Cronbach's Alpha or Kurder-Richardson's Coefficient Twenty is the Same for Two Tests', Psychometrika, Vol. 34, pp. 363-373, 1969.

GLESER, G.C., L.J. Cronbach, and N. Rajaratnam, 'Generalizability of Scores Influenced by Multiple Sources of Variance', Psychometrika, Vol. 30, pp. 395-418, 1965.

GULLIKSEN, H., Theory of Mental Tests, New York: John Wiley and Sons, 1950.

GUTTMAN, L., 'A Basis for Analyzing Test-Retest Reliability', Psychometrika, Vol. 10, pp. 255-282, 1945.

GUTTMAN, L., 'Reliability Formulas that do not Assume Experimental Independence', Psychometrika, Vol. 18, pp. 225-240, 1953.

HOYT, C., 'Test Reliability Estimated by Analysis of Variance', Psychometrika, Vol. 6, pp. 153-160, 1941.

HSU, Tse-Chi, and L. Feldt, 'The Effect of Limitations of the Number of Criterion Score Values on the Significance Level of the F-Test', American Educational Research Journal, Vol. 6, pp. 515-527, 1969.

IBM, Random Number Generating and Testing (C20-8011), New York: Technical Publication Department, IBM, 1959.

IBM, System/360 Scientific Subroutine Package (360A-CM-03X), Version III, Application Descriptions, 6th Ediction, New York: Technical Publication Department, IBM, 1968.

JACKSON, R.W., and G.A. Ferguson, Studies on the Reliability of Tests, Bulletin No. 12 of the Department of Educational Research, University of Toronto, 1941.

JÖRESKOG, K.G., 'Statistical Models for Congeneric Test Scores', 76th A. P. A. Proceedings, pp. 213-214, 1968.

JÖRESKOG, K.G., 'A General Method for Analysis of Covariance Structure', Biometrika, Vol. 57, pp. 239-251, 1970.

JÖRESKOG, K.G., 'Statistical Analysis of Sets of Congeneric Tests', Psychometrika, Vol. 36, pp. 109-133, 1971.

KEEPING, E.S., Introduction to Statistical Inference, Princeton: Van Nostrand, 1962.

KENDALL, M.G., and A. Stuart, The Advanced Theory of Statistics, Vol. 1, (1963); Vol. 2, (1967); Vol. 3, (1966), London: Charles Griffin.

KRISTOF, W., 'The Statistical Theory of Stepped-up Reliability Coefficient when a Test has been Divided into Several Equivalent Parts', Psychometrika, Vol. 28, pp. 221-238, 1963.

KRISTOF, W., 'Statistical Inferences about the Error Variance', Psychometrika, Vol. 28, pp. 129-143, 1963.

KRISTOF, W., 'Testing Differences Between Reliability Coefficients', The British Journal of Statistical Psychology, Vol. 17, pp. 105-111, 1964.

KRISTOF, W., 'Estimation of True Score and Error Variance for Tests Under Various Equivalence Assumptions', Psychometrika, Vol. 34, pp. 489-507, 1969.

KRISTOF, W., Statistical Notes on Reliability Estimation, Research Bulletin, RB-69-25, Educational Testing Service, 1969.

KRISTOF, W., 'On the Sampling Theory of Reliability Estimation', Journal of Mathematical Psychology, Vol. 7, pp. 371-377, 1970.

KUDER, G.F., and M.W. Richardson, 'The Theory of the Estimation of Test Reliability', Psychometrika, Vol. 2, pp. 151-160, 1937.

LaFORGE, R., 'Components of Reliability', Psychometrika, Vol. 30, pp. 187-195, 1965.

LEHMER, D.H., 'Mathematical Methods in Large-Scale Computing Units', Annals Computer Laboratory Harvard University, Vol. 16, pp. 141-146, 1951.

LORD, F., A Theory of Test Scores, Psychometric Monographs, No. 7, 1952.

LORD, F., 'Sampling Fluctuations Resulting from the Sampling of Test Items', Psychometrika, Vol. 20, pp. 1-22, 1955.

LORD, F., 'An Empirical Study of the Normality and Independence of Errors of Measurement in Test Scores', Psychometrika, Vol. 25, pp. 91-104, 1960.

LORD, F., 'Nominally and Rigorously Parallel Test Forms', Psychometrika, Vol. 29, pp. 335-345, 1964.

LORD, F., 'A Strong True-Score Theory with Application', Psychometrika, Vol. 30, pp. 239-250, 1965.

LORD, F., 'Estimating True-Score Distribution in Psychological Testing (An Empirical Bayes Estimation Problem)' Psychometrika, Vol. 34, pp. 259-299, 1969.

LORD, F., 'Item Characteristic Curves Estimated Without Knowledge of Their Mathematical Form: A Confrontation of Birnbaum's Logistic Model', Psychometrika, Vol. 35, pp. 43-50, 1970.

LORD, F., and M. Novick, Statistical Theories of Mental Test Scores, Reading, Massachusetts: Addison-Wesley, 1968.

MAGUIRE, T., and C. Hazlett, 'Reliability for the Researcher', Alberta Journal of Educational Research, Vol. 15, pp. 117-126, 1969.

MORRISON, D., Multivariate Statistical Methods, New York: McGraw-Hill, 1967.

NAYLOR, T., J. Balinfty, D. Burdick, and K. Chu, Computer Simulation Techniques, New York: John Wiley and Sons, 1968.

NITKO, A., The Power Functions of Some Proposed Tests for the Significance of Coefficient Alpha in the One-Sample and Two-Sample Cases, Ph.D. Thesis, University of Iowa, 1968.

NITKO, A., and L. Feldt, 'A Note on the Effect of Item Difficulty Distributions on the Sampling Distribution of KR20', American Educational Research Journal, Vol. 6, pp. 433-437, 1969.

NORTON, D., An Empirical Investigation of Some Effects of Non-Normality and Hetrogeneity on the F-Distribution, Ph.D. Thesis, State University of Iowa, 1952.

NOVICK, M., 'The Axioms and Principal Results of Classical Test Theory', Journal of Mathematical Psychology, Vol. 3, pp. 1-18, 1966.

NOVICK, M., and C. Lewis, 'Coefficient Alpha and the Reliability of Composite Measurements', Psychometrika, Vol. 32, pp. 1-13, 1967.

OLKIN, I., and J. Pratt, 'Unbiased Estimation of Certain Correlation Coefficients', Annals of Mathematical Statistics, Vol. 29, pp. 201-211, 1958.

PAYNE, W., and D. Anderson, 'Significance Levels for the Kuder-Richardson Twenty: An Automated Sampling Experiment Approach', Educational and Psychological Measurement, Vol. 28, pp. 23-39, 1968.

RAJARATNAM, N., 'Reliability Formulas for Independent Decision Data When Reliability Data are Matched', Psychometrika, Vol. 25, pp. 261-270, 1960

RAJARATNAM, N., L. Cronbach, and G. Gleser, 'Generalizability of Stratified-Parrallel Tests', Psychometrika, Vol. 30, pp. 39-56, 1965.

RAO, C., Linear Statistical Inference and Its Applications, New York: John Wiley and Sons, 1965.

ROZEBOOM, W.W., Foundations of the Theory of Prediction, Homewood, Illinois: Dorsey Press, 1966.

SAMEJIMA, F., Estimation of Latent Ability Using a Response Pattern of Graded Scores, No. 17, Psychometrika Monograph Supplement, 1969.

SCHEFFÉ, H., The Analysis of Variance, New York: John Wiley and Sons, 1959.

SHOEMAKER, D., An Empirical Study of Generalizability Coefficients for Matched and Unmatched Data, Ph.D. Thesis, University of Houston, 1966.

SPEARMAN, C., 'Correlation Calculated from Faulty Data', British Journal of Psychology, Vol. 3, pp. 271-295, 1910.

SRIVASTAVA, A., and H. Webster, 'An Estimation of True Scores in the Case of Items Scored on a Continuous Scale', Psychometrika, Vol. 32, pp. 327-338, 1967.

STUDENT, 'The Probable Error of Means', Biometrika, Vol. 6, pp. 1-25, 1908.

TEICHROEW, D., 'History of Distribution Sampling Priori to the Era of the Computer and its Relevance to Simulation', Journal of American Statistical Association, Vol. 60, pp. 27-49, 1965.

TOCHER, K.D., 'The Application of Automatic Computers to Sampling Experiments', Journal of the Royal Statistical Society, B16, pp. 39-61, 1954.

TRYON, R., 'Reliability and Behavior Domain Validity: Reformulation and Historical Critique', Psychological Bulletin, Vol. 54, pp. 229-249, 1957.

TUKEY, J.W., 'Variance of Variance Components: 1. Balanced Design', Annals of Mathematical Statistics, Vol. 27, pp. 722-736, 1956.

VOTAW, D.F., Jr., 'Testing Compound Symmetry in a Normal Multi-Variate Distribution, Annals of Mathematical Statistics, Vol. 19, pp. 447-473, 1948.

WEITZMAN, R., 'Monte Carlo Studies of a Single-Trial Estimation of the Test-Retest Reliability of a Multiple-Choice Test', Proceedings, 77th APA Convention, Part I, pp. 121-122, 1969.

WINER, B., Statistical Principles in Experimental Design, New York: McGraw-Hill, 1962.

WOODBURY, M., 'The Stochastic Model of Mental Testing Theory and an Application', Psychometrika, Vol. 28, pp. 391-393, 1963.

APPENDIX A.1

LISTINGS OF COMPUTER PROGRAMS

REL01 : Simulation Program for Continuous Part Test Case

REL02 : Simulation Program for Binary Item Test Case

REL00 : A Package of Sub-Programs Shared by REL01 and REL02

```
FORTRAN IV G COMPILER         MAIN             09-15-71          15:55.55        PAGE 0001
           C
           C   REL01          DIVISION OF EDUCATIONAL RESEARCH SERVICES
           C                  UNIVERSITY OF ALBERTA
           C       ****************************************************************************
           C   PURPOSE:       SIMULATES CONTINUOUS PART-TEST SCORES BASED ON MULTI-
           C                  FACTOR TRUE SCORE MODEL TO INVESTIGATE SAMPLING
           C                  DISTRIBUTION OF RELIABILITY ESTIMATES
           C   CARD INPUT:
           C                  1. TITLE(20A4)
           C                  2. PARAMETERS(11I5,F5.5):NSAM,MI,MJ,NF,IX,IDIST,IDISE,
           C                     IPUNCH,IPLOT,MODE,LB,SIG
           C                     NSAM     NO OF SAMPLES SIMULATED
           C                     MI       NO OF SUBJECTS IN THE SAMPLE
           C                     MJ       NO OF PARTS
           C                     NF       NO OF FACTORS IN TRUE SCORE
           C                     IX       ANY ODD INTEGER TO INITIATE RANDOM NUMBER
           C                     IDIST    OPTION FOR THE DISTRIBUTION OF RANDOM
           C                              EFFECTS (TRUE SCORE)
           C                              0-NORMAL
           C                              1-SPECIFIED BY SUBPROGRAM DIST
           C                     IDISE    OPTION FOR THE DISTRIBUTION OF ERROR
           C                              0-NORMAL
           C                              1-SPECIFIED BY SUBPROGRAM DISE
           C                     IPUNCH   OPTION FOR CARD OUT PUT OF FREQUENCIES
           C                              0-NO CARD OUTPUTS
           C                              1-CARD OUTPUT REQUIRED
           C                     IPLOT    OPTION FOR PLOTS
           C                              0-NOT REQUIRED
           C                              1-REQUIRED
           C                     MODE     OPTION FOR ESTIMATION FORMULA
           C                              0-ALPHA FORMULA(BIASED)
           C                              1-KRISTOF CORRECTION(UNBIASED)
           C                              2-BOTH OF ABOVE
           C                     LB       OPTION FOR THE NO OF CLASS INTERVALS FOR
           C                              THE FREQUENCY CALCULATION,24,36 OR 48,
           C                              ASSUMED 24
           C                     SIG      SIGNIFICANCE LEVEL FOR EACH TAIL,ASSUMED
           C                              0.05
           C                  3. FMT      FORMAT FOR THE INPUT VECTORS AND MATRIX
           C                  4. FIX      A VECTOR OF MEANS FOR EACH PART
           C                  5. ERR      A VECTOR OF STANDARD DEVIATION OF ERROR
           C                              SCORES FOR EACH PART-TEST
           C                  6. FAC      A FACTOR LOADING MATRIX OF SIZE MJ BY NF
           C                  7. A BLANK CARD
           C   REMARK:
           C                   1.CURRENTLY DIMENSIONED TO ACCOMODATE UP TO FOLLOWING
           C                     SIZE PARAMETERS
           C                      NSAM                                          5000
           C                      MI                                             100
           C                      MJ                                              30
           C                      NF                                              10
           C                      LB                                              48
           C   SUBPROGRAMS:   (FORTRAN) ANOV,BOXSN,CHIPRB,COUNT,DISCRP,DISP,EXAMPL,
           C                  FISHER,FITTES,FST,MXOUT,PLOT,POPR,PUNCH,RELDIS,ROZB,
           C                  SIGTES,VARXX,VECRAN,VEOUT,  DATA,DIST,DISE
```

```
FORTRAN IV G COMPILER        MAIN              09-15-71           15:55.55        PAGE 0002
                C                     (*SSPLIB) BDTR,CDTR,DLGAM,NDTR,RANK
                C     PROGRAMMER:     K.BAY
                C       NMAX=LARGER OF NSAM AND MI*MJ
                C       DIMENSION FAC(MJ*NF),ERR(MJ),VAR(MJ),DIS(MJ,MJ),FIX(MJ),BSUM(MJ),
                C     1BSS(MJ),Y(MI*MJ),BB(MJ),FMS(NSAM*3),FREQ(LB*3),TEMP(MI*NF)
 0001                   DIMENSION TITLE(20),FMT(20),LAB( 4),FAC( 300),ERR( 30),VAR( 30),
                      1DIS( 900),FIX( 30),BSUM( 30),BSS( 30),Y( 3000),BB( 30),FMS( 15000)
                      2,FREQ(144),XBAR( 6),XVAR( 6),TEMP(1000)
 0002                   REAL*8 LAB
 0003                   DATA LAB/'SUBJECT','PARTTEST','ERROR','REL COF'/
 0004               100 FORMAT(20A4)
 0005               101 FORMAT(1H1,20A4)
 0006               102 FORMAT(11I5,F5.5)
 0007               103 FORMAT(/,10X,'NO OF SAMPLES SIMULATED',15X, I4,/,10X,'NO OF SUBJEC
                      1TS IN EACH SAMPLE',11X,I2,/,10X,'NO OF PART-TESTS',24X,I2,/,10X,
                      2'NO OF FACTORS IN TRUE SCORE',14X,I1,/,10X,'STARTING INTEGER RANDO
                      3M NUMBER', 2X,I10,/,10X,'OPTION FOR CARD OUTPUT',19X,I1,/,10X,
                      4'OPTION FOR PLOT',26X,I1,/,10X,'OPTION FOR ESTIMATION FORMULA',
                      512X,I1,/,10X,'OPTION FOR THE NO OF CLASS INTERVALS',3X,I3,/,10X,
                      6'SIGNIFICANCE LEVEL',19X,F5.3,/)
 0008               104 FORMAT(2X,I3,5X,2E14.6,2X,'|',2X,2E14.6)
 0009               105 FORMAT(/,1X,'*******LAST SEED RANDOM NUMBER  IX=',I10)
 0010               106 FORMAT(/,1X,'DISCRIPTIVE STATISTICS FOR FIXED EFFECT ESTIMATES AND
                      1 EXPECTED VALUES UNDER M.F.  MODEL',/,1X,'PART',7X,'MEAN',10X,'EXP
                      2ECTED',6X,'|',5X,'VARIANCE',8X,'EXPECTED')
 0011               107 FORMAT(1H1,23('@'),/,1X,'@',2X,'SUMMARY OF OUTPUT',2X,'@',/,1X,23(
                      1'@'))
 0012               108 FORMAT(10X,'ERROR SCORE DISTRIBUTIONS ARE NORMAL')
 0013               109 FORMAT(10X,'ERROR SCORE DISTRIBUTIONS ARE NOT NORMAL')
 0014               110 FORMAT(10X,'TRUE SCORE DISTRIBUTIONS ARE NORMAL')
 0015               111 FORMAT(10X,'TRUE SCORE DISTRIBUTIONS ARE NOT NORMAL')
 0016                10 READ(5,100) TITLE
 0017                   IF(TITLE(1).EQ.TITLE(2)) GO TO 99
 0018                   WRITE(6,101) TITLE
 0019                   READ(5,102) NSAM,MI,MJ,NF,IX,IDIST,IDISE,IPUNCH,IPLOT,MODE,LB,SIG
 0020                   IF(SIGLL.LE.0.0) SIG=0.05
 0021                   IF(LB.NE.24.AND.LB.NE.36.AND.LB.NE.48) LB=24
 0022                   IF(NF.LE.0) NF=1
 0023                   WRITE(6,103) NSAM,MI,MJ,NF,IX,IPUNCH,IPLOT,MODE,LB,SIG
 0024                   IF(IDIST.EQ.0) WRITE(6,110)
 0025                   IF(IDIST.EQ.1) WRITE(6,111)
 0026                   IF(IDISE.EQ.0) WRITE(6,108)
 0027                   IF(IDISE.EQ.1) WRITE(6,109)
 0028                   CALL PCPR(MJ,NF,FAC,ERR,VAR,DIS,REL,ALPHA,TVAR,EVAR,FMT,FIX,GMEAN)
 0029                   TVAR=TVAR/(MJ*MJ)
 0030                   EVAR=EVAR/MJ
 0031                   THETA=TVAR/EVAR
 0032                   DIV=1.0+MJ*THETA
 0033                   DO 20 J=1,MJ
 0034                   BSUM(J)=0.0
 0035                 20 BSS(J)=0.0
 0036                   CALL VECRAN(Y,300,IX)
 0037                   CALL EXAMPL(MI,MJ,NF,FAC,ERR,IDIST,IDISE,IX,Y,FIX,TEMP,BB,FMS,REL)
 0038                   DO 50 NTRIAL=1,NSAM
 0039                   CALL DATA(MI,MJ,NF,FAC,ERR,IDIST,IDISE,IX,Y,FIX,TEMP)
```

```
FORTRAN IV G COMPILER          MAIN              09-15-71           15:55.55          PAGE 0003
                CALL ANOV(Y,MI,MJ,FMSA,FMSB,FMSE,BB)
                FMS(NTRIAL)=FMSA
                II=NTRIAL+NSAM
                FMS(II)=FMSB
                II=II+NSAM
                FMS(II)=FMSE
                DO 45 J=1,MJ
                BSUM(J)=BSUM(J)+BB(J)
             45 BSS(J)=BSS(J)+BB(J)**2
             50 CONTINUE
                DO 54 J=1,MJ
             54 FIX(J)=FIX(J)-GMEAN
                WRITE(6,107)
          C     CALL MXOUT(FMS,NSAM,3,0,44,44HMEAN SQUARES: COL-1 MSA,COL-2 MSB,CO
                NN=NSAM*3
                XMAX=0.0
                DO 56 I=1,NN
             56 IF(FMS(I).GT.XMAX) XMAX=FMS(I)
                NN=XMAX/10.0+1.0
                XMAX=NN*10.0
                TINT=XMAX/LB
                CALL COUNT(FMS,NSAM,3,0.0,TINT,LB,FREQ,XBAR,XVAR)
                XMIN=0.0
                XMAX=TINT*LB
                XBAR(4)=MJ*TVAR+EVAR
                XXXX=0.0
                DO 58 J=1,MJ
             58 XXXX=XXXX+FIX(J)**2
                DFA=MI-1
                DFB=MJ-1
                DFE=(MI-1)*(MJ-1)
                XXXX=XXXX/DFB
                XBAR(5)=EVAR+XXXX*MI
                XBAR(6)=EVAR
                XVAR(4)=(2.0*(MJ*TVAR+EVAR)**2)/DFA
                XVAR(5)=((EVAR+2.0*MI*XXXX    )*2.0*EVAR)/DFB
                XVAR(6)=(2.0*EVAR*EVAR)/DFE
                CALL DISCRP(3,XBAR,XVAR,LAB(1),52,52H MEAN SQUARES AND EXPECTED VA
               1LUES UNDER ANOVA MODEL )
                CALL VARXX(NSAM,MJ,BSUM,BSS)
                WRITE(6,106)
                DO 63 J=1,MJ
                XX=0.0
                DO 61 M=1,MJ
                D1=-1.0/MJ
                IF(J.EQ.M) D1=D1+1.0
                DO 60 K=1,MJ
                D2=-1.0/MJ
                IF(J.EQ.K) D2=D2+1.0
                MK=MJ*(M-1)+K
             60 XX=XX+D1*D2*DIS(MK)
             61 CONTINUE
                XX=XX/MI
             63 WRITE(6,104) J,BSUM(J),FIX(J),BSS(J),XX
                II=NSAM*2
```

FORTRAN IV G COMPILER MAIN 09-15-71 15:55.55 PAGE 0004

```
0093            DO 70 I=1,NSAM
0094            II=II+1
0095         70 FMS(I)=1.0-FMS(II)/FMS(I)
0096            CALL RANK(FMS(1),FMS(NSAM+1),NSAM)
0097            IF(MODE.NE.1) CALL RELDIS(FMS,NSAM,DFA,DFE,FREQ,LB,REL,TEMP,SIG,
               1XBAR,XVAR,IPUNCH,IPLOT,0,LAB)
0098            IF(MODE.NE.0) CALL RELDIS(FMS,NSAM,DFA,DFE,FREQ,LB,REL,TEMP,SIG,
               1XBAR,XVAR,IPUNCH,IPLOT,1,LAB)
0099            WRITE(6,105) IX
0100            GO TO 10
0101         99 STOP
0102            END
```

TOTAL MEMCRY REQUIREMENTS 0152B4 BYTES

```
FORTRAN IV G COMPILER        DATA              09-15-71          15:56.00        PAGE 0001

 0001                 SUBROUTINE DATA(MI,MJ,NF,FAC,ERR,IDIST,IDISE,IX,Y,FIX,TEMP)
              C    PURPOSE        CREATES DATA MATRIX FOR REL01
              C          MI       SAMPLE SIZE
              C          MJ       NO OF PARTS
              C          NF       NO OF FACTORS IN TRUE SCORE
              C          FAC      INPUT FACTOR LOADING MATRIX
              C          ERR      INPUT VECTOR OF STANDARD DEVIATION FOR ERRORS
              C          IDIST    OPTION FOR TRUE SCORE DISTRIBUTION
              C          IDISE    OPTION FOR ERROR SCORE DISTRIBUTION
              C          IX       SEED ODD INTEGER RANDOM NUMBER
              C          FIX      INPUT FIXED EFFECT VECTOR
              C          TEMP     WORKING MATRIX
 0002                 DIMENSION FAC(MJ,NF),ERR(MJ),Y(MI,MJ),FIX(MJ),TEMP(MI,NF)
              C       IF(IDISE.EQ.0) CALL SRAND(Y,(MI*MJ),IX)
 0003                 IF(IDISE.EQ.0) CALL BOXSN(Y,(MI*MJ),IX)
 0004                 IF(IDISE.EQ.1) CALL DISE(Y,MI,MJ,IX)
 0005                 DO 20 I=1,MI
 0006                 DO 20 J=1,MJ
 0007              20 Y(I,J)=FIX(J)+Y(I,J)*ERR(J)
              C       IF(IDIST.EQ.0) CALL SRAND(TEMP,(MI*NF),IX)
 0008                 IF(IDIST.EQ.0) CALL BOXSN(TEMP,(MI*NF),IX)
 0009                 IF(IDIST.EQ.1) CALL DIST(TEMP,MI,NF,IX)
 0010                 DO 30 I=1,MI
 0011                 DO 25 K=1,NF
 0012                 DO 23 J=1,MJ
 0013              23 Y(I,J)=Y(I,J)+FAC(J,K)*TEMP(I,K)
 0014              25 CONTINUE
 0015              30 CONTINUE
 0016                 RETURN
 0017                 END

  TOTAL MEMORY REQUIREMENTS 00050E BYTES
```

```
FORTRAN IV G COMPILER        EXAMPL          09-15-71         15:56.02        PAGE 0001

0001            SUBROUTINE EXAMPL(MI,MJ,NF,FAC,ERR,IDIST,IDISE,IX,Y,FIX,TEMP,BB,
               1FMS,REL)
          C   PURPOSE         GIVES EXAMPLE OUTPUTS FOR REL01
          C       ARGUMENTS THE SAME AS THE MAINLINE PROGRAM REL01
          C   SUBPROGRAM      DATA,MXOUT,ANOV,VEOUT,DISP,ROZ8
0002            DIMENSION FAC(MJ,NF),ERR(MJ),Y(MI,MJ),FIX(MJ),TEMP(MI,NF),BB(MJ),
               1FMS(MJ,MJ)
0003        100 FORMAT(1H1,20('@'),/,1X,'@',3X,'EXAMPLE RUNS',3X,'@',/,1X,20('@'))
0004        101 FORMAT(/,1X,'MSA=',E14.6,2X,'MSB=',E14.6,2X,'MSE=',E14.6,2X,'F=',
               1E14.6,2X,'ALPHA=',F8.5,2X,'UNBIASED REL EST(ANOVA)=',F8.5)
0005        102 FORMAT(/,1X,'SAMPLE DISPERSION : SATURATION COEFF=',F9.5,3X,'HOMOG
               1ENEITY COEFF=',F9.5,5X,'HOM/SAT=',F9.5)
0006        103 FORMAT(/,1X,'GMEAN=',E14.6)
0007        104 FORMAT(/,1X,'VARIANCE OF ALPHA ESTIMATE UNDER ANOVA=',F8.5,3X,
               1'ESTIMATE=',F8.5)
0008        105 FORMAT(/,1X,'VARIANCE OF UNBIASED REL ESTIMATES UNDER ANOVA=',F8.5
               1,3X,'ESTIMATE=',F8.5)
0009            WRITE(6,100)
0010            CALL DATA(MI,MJ,NF,FAC,ERR,IDIST,IDISE,IX,Y,FIX,TEMP)
0011            CALL MXOUT(Y,MI,MJ,0,12,12HDATA MATRIX )
0012            CALL ANOV(Y,MI,MJ,FMSA,FMSB,FMSE,BB)
0013            FF=FMSA/FMSE
0014            AL=1.0-1.0/FF
0015            ALL=(2.0+(MI-3.0)*AL)/(MI-1.0)
0016            WRITE(6,101) FMSA,FMSB,FMSE,FF,AL,ALL
0017            CALL VEOUT(BB,MJ,28,28HSAMPLE FIXED EFFECTS VECTOR )
0018            CALL DISP(Y,MI,MJ,FMS,BB)
0019            CALL VEOUT(BB,MJ,20,20HSAMPLE MEANS VECTOR )
0020            SUM=0.0
0021            DO 20 J=1,MJ
0022         20 SUM=SUM+BB(J)
0023            SUM=SUM/MJ
0024            WRITE(6,103) SUM
0025            CALL MXOUT(FMS,MJ,MJ,0,24,24HSAMPLE DISPERSION MATRIX)
0026            CALL ROZB(FMS,MJ,SAT,HOM)
0027            ALPHA=HOM/SAT
0028            WRITE(6,102) SAT,HOM,ALPHA
0029            IF(MI.LE.5) GO TO 90
0030            VF=(2.0*(MI-1.0)*(MJ*MI-MJ-2.0))/((MI-5.0)*(MJ-1.0)*(MI-3.0)**2)
0031            VARA=VF*(1-REL)**2
0032            VARE=VF*(1-AL)**2
0033            WRITE(6,104) VARA,VARE
0034            C2=(MI-3.0)/(MI-1.0)
0035            VARA=VARA*C2*C2
0036            VARE=VARE*C2*C2
0037            WRITE(6,105) VARA,VARE
0038         90 RETURN
0039            END

 TOTAL MEMORY REQUIREMENTS 000A04 BYTES
```

```
FORTRAN IV G COMPILER        POPR            09-15-71          15:56.04          PAGE 0001

0001                SUBROUTINE POPR(MJ,NF,FAC,ERR,VAR,DIS,REL,ALPHA,TVAR,EVAR,FMT,FIX,
                   1GMEAN)
              C    PURPOSE            PERFORMS BASIC COMPUTATIONS FOR POPULATION PARAMETERS
              C          MJ           NO OF PART-TESTS
              C          NF           NO OF FACTORS IN TRUE SCORE
              C          FAC          FACTOR LOADING MATRIX
              C          ERR          ERROR STANDARD DEVIATION VECTOR
              C          VAR          OUTPUT VECTOR FOR VARIANCES OF PARTS
              C          DIS          OUTPUT DISPERSION MATRIX OF PART-SCORE VECTOR
              C          REL          OUTPUT POPULATION RELIABILITY
              C          ALPHA        OUTPUT POPULATION RELIABILITY
              C          TVAR         OUTPUT TRUE SCORE VARIANCE OF TEST SCORE
              C          EVAR         OUTPUT ERROR SCORE VARIANCE OF TEST SCORE
              C          FMT          FORMAT FOR INPUT VECTORS AND MATRIX
              C          FIX          OUTPUT MEAN VECTOR FOR PARTS
              C          GMEAN        OUTPUT GENERAL MEAN
0002                DIMENSION FAC(MJ,NF),ERR(MJ),VAR(MJ),DIS(MJ,MJ),FMT(20),FIX(MJ)
0003            100 FORMAT(/,1X,'POPULATION PARAMETERS',/,1X,'RELIABILITY',19X,F9.5,/,
                   11X,'ALPHA',25X,F9.5,/,1X,'TRUE SCORE VARIANCE',11X,E14.6,/,1X,'ERR
                   2OR SCORE VARIANCE',10X,E14.6)
0004            101 FORMAT(/,1X,'TRUE SCORE DISPERSION: SATURATION COEFF=',F9.5,5X,'HO
                   1MOGENEITY COEFF=',F9.5)
0005            102 FORMAT(1H1,37('@'),/,1X,'@',3X,'INPUT POPULATION PARAMETERS',5X,
                   1'@',/,1X,37('@'))
0006            103 FORMAT(/,1X,'GMEAN=',E14.6)
0007            104 FORMAT(20A4)
0008            105 FORMAT(/,/,1X,'FORMAT FOR THE DATA',5X,20A4)
0009            106 FORMAT(/,1X,'PART SCORE DISPERSION: SATURATION COEFF=',F9.5,5X,'HO
                   1MOGENEITY COEF=',F9.5)
0010                WRITE(6,102)
0011                READ(5,104) (FMT(I),I=1,20)
0012                WRITE(6,105) (FMT(I),I=1,20)
0013                READ(5,FMT) (FIX(J),J=1,MJ)
0014                READ(5,FMT)(ERR(J),J=1,MJ)
0015                DO 10 I=1,MJ
0016             10 READ(5,FMT)(FAC(I,J),J=1,NF)
0017                CALL VEOUT(FIX,MJ,12,12HMEANS VECTOR)
0018                CALL VEOUT(ERR,MJ,32,32HERROR STANDARD DEVIATIONS VECTOR)
0019                CALL MXOUT(FAC,MJ,NF,0,24,24HFACTOR LOADING MATRIX    )
0020                TVAR=0.0
0021                EVAR=0.0
0022                DO 15 J=1,MJ
0023                VAR(J)=0.0
0024                DO 12 I=1,MJ
0025                DIS(I,J)=0.0
0026                DO 11 K=1,NF
0027             11 DIS(I,J)=DIS(I,J)+FAC(I,K)*FAC(J,K)
0028                TVAR=TVAR+DIS(I,J)
0029             12 CONTINUE
0030                VAR(J)=DIS(J,J)+ERR(J)*ERR(J)
0031             15 EVAR=EVAR+ERR(J)*ERR(J)
0032                CALL ROZB(DIS,MJ,SAT,HOM)
0033                CALL MXOUT(DIS,MJ,MJ,0,28,28HTRUE SCORE DISPERSION MATRIX)
0034                WRITE(6,101) SAT,HOM
0035                REL=TVAR/(TVAR+EVAR)
```

```
          DO 16 J=1,MJ
       16 DIS(J,J)=VAR(J)
          CALL ROZB(DIS,MJ,SAT,HOM)
          CALL MXOUT(DIS,MJ,MJ,0,20,20HDISPERSION MATRIX    )
          WRITE(6,106) SAT,HCM
          ALPHA=HOM/SAT
          WRITE(6,100) REL,ALPHA,TVAR,EVAR
          GMEAN=0.0
          DO 20 J=1,MJ
       20 GMEAN=GMEAN+FIX(J)
          GMEAN=GMEAN/MJ
          CO 21 J=1,MJ
       21 FIX(J)=FIX(J)-GMEAN
          WRITE(6,103) GMEAN
          CALL VEOUT(FIX,MJ,16,16HFIXED EFFECTS    )
          CALL VEOUT(VAR,MJ,20,20HVARIANCES OF PARTS   )
          DO 23 J=1,MJ
       23 FIX(J)=FIX(J)+GMEAN
          RETURN
          END
```

TCTAL MEMCRY REQUIREMENTS 000BDC BYTES

```
FORTRAN IV G COMPILER        DIST             09-15-71         15:56.10        PAGE 0001

 0001               SUBROUTINE DIST(TEMP,MI,NF,IX)
              C    PURPOSE         CREATE STANDARD RANDOM TRUE SCORE MATRIX FOR REL01
              C         TEMP       OUTPUT TRUE SCORE MATRIX
              C         MI         NO OF ROWS OF TEMP
              C         NF         NO OF COLS OF TEMP
              C         IX         SEED ODD INTEGER RANDOM NUMBER
              C****THIS EXAMPLE PRODUCES EXPONENTIAL TRUE SCORES
 0002               DIMENSION TEMP(MI,NF)
 0003               CALL VECRAN(TEMP,(MI*NF),IX)
 0004               DO 20 I=1,MI
 0005               DO 10 J=1,NF
 0006            10 TEMP(I,J)=-ALOG(TEMP(I,J))-1.0
 0007            20 CONTINUE
 0008               RETURN
 0009               END

  TOTAL MEMORY REQUIREMENTS 00025A BYTES
```

```
FORTRAN IV G COMPILER        DISE              09-15-71          15:56.11        PAGE 0001

 0001                 SUBROUTINE DISE(Y,MI,MJ,IX)
               C   PURPOSE        CREATE STANDARD RANDOM ERROR MATRIX Y FOR REL01
               C        Y         OUTPUT MATRIX
               C        MI        NO OF ROWS OF Y
               C        MJ        NO OF COLS OF Y
               C        IX        SEED ODD INTEGER RANDOM NUMBER
               C****THIS EXAMPLE PRODUCES UNIFORM ERROR SCORES
 0002                 DIMENSION Y(MI,MJ)
 0003                 CALL VECRAN(Y,(MI*MJ),IX)
 0004                 SQR=SQRT(12.0)
 0005                 DO 20 J=1,MJ
 0006                 DO 10 I=1,MI
 0007              10 Y(I,J)=(Y(I,J)-0.5)*SQR
 0008              20 CONTINUE
 0009                 RETURN
 0010                 END

  TOTAL MEMORY REQUIREMENTS 000264 BYTES
5:56.11 10.5 RC=0
```

```
C
C    REL02           DIVISION OF EDUCATIONAL RESEARCH SERVICES
C                    UNIVERSITY OF ALBERTA
C         *************************************************************
C    PURPOSE         SIMULATE UNIT-DICHOTOMOUS(BINARY) ITEM TEST SCORES
C                    BASED ON NORMAL OGIVE MODEL TO INVESTIGATE SAMPLING
C                    DISTRIBUTION OF RELIABILITY ESTIMATES
C    CARD INPUT:
C                    1. TITLE(20A4)
C                    2. PARAMETERS(10I5,F5.5):NSAM,MI,MJ,IX,IDIST,IDISE,
C                       IPUNCH,IPLOT,MODE,LB,SIG
C                       NSAM     NO OF SAMPLES SIMULATED
C                       MI       NO OF SUBJECTS IN THE SAMPLE
C                       MJ       NO OF ITEMS
C                       IX       ANY ODD INTEGER TO INITIATE RANDOM NUMBER
C                       IDIST    OPTION FOR THE DISTRIBUTION OF RANDOM
C                                EFFECTS (TRUE SCORE)
C                                0-NORMAL
C                                1-SPECIFIED BY SUBPROGRAM DIST
C                       IDISE    OPTION FOR THE DISTRIBUTION OF ERROR
C                                0-NORMAL
C                                1-SPECIFIED BY SUBPROGRAM DISE
C                       IPUNCH   OPTION FOR CARD OUT PUT OF FREQUENCIES
C                                0-NO CARD OUTPUTS
C                                1-CARD OUTPUT REQUIRED
C                       IPLOT    OPTION FOR PLOTS
C                                0-NOT REQUIRED
C                                1-REQUIRED
C                       MODE     OPTION FOR ESTIMATION FORMULA
C                                0-ALPHA FORMULA(BIASED)
C                                1-KRISTOF CORRECTION(UNBIASED)
C                                2-BOTH OF ABOVE
C                       LB       OPTION FOR THE NO OF CLASS INTERVALS FOR
C                                FREQUENCY CALCULATION,24,36 OR 48,ASSMED 24
C                       SIG      SIGNIFICANCE LEVEL FOR EACH TAIL,ASSUMED
C                                0.05
C                    3. FMT      FORMAT FOR THE INPUT VECTORS
C                    4. DIF      A VECTOR OF ITEM DIFFICULTIES
C                    5. BS       A VECTOR OF BISERIAL CORRELATIONS
C                    6. A BLANK CARD
C    REMARK:
C                     1.CURRENTLY DIMENSIONED TO ACCOMODATE UP TO FOLLOWING
C                       SIZE PARAMETERS
C                        NSAM                                        5000
C                        MI                                           100
C                        MJ                                            30
C                        LB                                            48
C    SUBPROGRAMS:    (FORTRAN) ANOV,BOXSN,CHIPRB,COUNT,DISCRP,DISP,EXAMPL,
C                    FISHER,FITTES,FST,ITEMCO,MXOUT,PARALL,PLOT,POPR,PUNCH,
C                    RELDIS,ROZB,SIGTES,VARXX,VECRAN,VEOUT,DATA,DIST,DISE
C                    (*SSPLIB) BDTR,CDTR,DLGAM,NDTR,NDTRI,RANK
C    PROGRAMMER:      K.BAY
C      NMAX=LARGER OF NSAM AND MI*MJ
C      DIMENSION BS(MJ),ERR(MJ),VAR(MJ),DIS(MJ*MJ),DIF(MJ),BSUM(MJ),
C     1BSS(MJ),Y(MI*MJ),BB(MJ),FMS(NSAM*3),FREQ(LB*3),TEMP(MI),RR(MJ),
```

```
               C      R(MJ*(MJ+1)/2),XBAR(2*MJ),XVAR(MJ)
 0001                 DIMENSION TITLE(20),FMT(20),LAB( 4),BS( 30),ERR( 30),VAR( 30),
                     1DIS( 900),DIF( 30),BSUM( 30),BSS( 30),Y( 3000),BB( 30),FMS( 15000)
                     2,FREQ(144),XBAR(60),XVAR(30),TEMP( 100),RR( 30),PP( 30),R(465),
                     3X( 3000)
 0002                 REAL*8 LAB
 0003                 DATA LAB/'SUBJECT','ITEMS   ','ERROR','REL COF'/
 0004             100 FORMAT(20A4)
 0005             101 FORMAT(1H1,20A4)
 0006             102 FORMAT(10I5,F5.5)
 0007             103 FORMAT(/,10X,'NO OF SAMPLES SIMULATED',15X, I4,/,10X,'NO OF SUBJEC
                     1TS IN EACH SAMPLE',10X,I3,/,10X,'NO OF ITEMS',28X,I3,/,10X,'STARTI
                     2NG SEED RANDOM NUMBER', 5X,I10,/,10X,'OPTION FOR CARD OUTPUT',19X,
                     3I1,/,10X,'OPTION FOR PLOT',26X,I1,/,10X,'OPTION FOR ESTIMATION FOR
                     4MULA',12X,I1,/,10X,'OPTIION FOR CLASS INTERVALS',12X,I3,/,10X,
                     5'SIGNIFICACNCE LEVEL',19X,F5.3,/)
 0008             104 FORMAT(2X,I3,5X,2E14.6,2X,'|',2X,2E14.6)
 0009             105 FORMAT(/,1X,'*******LAST SEED RANDOM NUMBER  IX=',I10)
 0010             106 FORMAT(/,1X,'DISCRIPTIVE STATISTICS FOR FIXED EFFECT ESTIMATES AND
                     1 EXPECTED VALUES UNDER M.F.  MODEL',/,1X,'PART',7X,'MEAN',10X,'EXP
                     2ECTED',6X,'|',5X,'VARIANCE',8X,'EXPECTED')
 0011             107 FORMAT(1H1,23('@'),/,1X,'@',2X,'SUMMARY OF OUTPUT',2X,'@',/,1X,23(
                     1'@'))
 0012             108 FORMAT(10X,'ERROR SCORE DISTRIBUTIONS ARE NORMAL')
 0013             109 FORMAT(10X,'ERROR SCORE DISTRIBUTIONS ARE NOT NORMAL')
 0014             110 FORMAT(10X,'TRUE SCORE DISTRIBUTIONS ARE NORMAL')
 0015             111 FORMAT(10X,'TRUE SCORE DISTRIBUTIONS ARE NOT NORMAL')
 0016              10 READ(5,100) TITLE
 0017                 IF(TITLE(1).EQ.TITLE(2)) GO TO 99
 0018                 WRITE(6,101) TITLE
 0019                 READ(5,102) NSAM,MI,MJ,IX,IDIST,IDISE,IPUNCH,IPLOT,MODE,LB,SIG
 0020                 IF(SIGLL.LE.0.0) SIG=0.05
 0021                 IF(LB.NE.24.AND.LB.NE.36.AND.LB.NE.48) LB=24
 0022                 WRITE(6,103) NSAM,MI,MJ,IX,IPUNCH,IPLOT,MODE,LB,SIG
 0023                 IF(IDIST.EQ.0) WRITE(6,110)
 0024                 IF(IDIST.EQ.1) WRITE(6,111)
 0025                 IF(IDISE.EQ.0) WRITE(6,108)
 0026                 IF(IDISE.EQ.1) WRITE(6,109)
 0027                 CALL POPR(MJ,BS,DIF,R,REL,ALPHA,FMT,RR,PP,ERR,TEMP,TVAR,EVAR,DIS)
 0028                 IR=0.0
 0029                 DO 20 J=1,MJ
 0030                 BSUM(J)=0.0
 0031                 XVAR(J)=0.0
 0032                 XBAR(J)=0.0
 0033                 JJ=MJ+J
 0034                 XBAR(JJ)=0.0
 0035                 BSS(J)=0.0
 0036                 DO 19 I=1,J
 0037                 IR=IR+1
 0038              19 R(IR)=0.0
 0039              20 CONTINUE
 0040                 CALL VECRAN(Y,100,IX)
 0041                 CALL EXAMPL(MI,MJ,BS,ERR,TEMP,RR,Y,IX,IDIST,IDISE,X,XBAR,BB,FMS,
                     1REL,R,XVAR)
 0042                 DO 50 NTRIAL=1,NSAM
```

```
FORTRAN IV G COMPILER        MAIN            09-15-71         15:54.53        PAGE 0003
      CALL DATA(MI,MJ,BS,ERR,RR,Y,X,TEMP,IX,IDIST,IDISE,XBAR,R,XVAR)
      CALL ANOV(Y,MI,MJ,FMSA,FMSB,FMSE,BB)
      FMS(NTRIAL)=FMSA
      II=NTRIAL+NSAM
      FMS(II)=FMSB
      II=II+NSAM
      FMS(II)=FMSE
      DO 45 J=1,MJ
      BSUM(J)=BSUM(J)+BB(J)
   45 BSS(J)=BSS(J)+BB(J)**2
   50 CONTINUE
C     CALL MXOUT(FMS,NSAM,3,0,44,44HMEAN SQUARES: COL-1 MSA,COL-2 MSB,CO
      CALL PARALL(R,XBAR,MI,MJ,NSAM,XVAR,TVAR2,EVAR2,REL2)
      WRITE(6,107)
      IF(IDIST.EQ.0.AND.ISISE.EQ.0) GO TO 53
      TVAR=TVAR2
      EVAR=EVAR2
      REL=REL2
      DO 51 J=1,MJ
   51 DIF(J)=XBAR(J)
   53 TVAR=TVAR/(MJ*MJ)
      EVAR=EVAR/MJ
      THETA=TVAR/EVAR
      DIV=1.0+MJ*THETA
      GMEAN=0.0
      DO 54 J=1,MJ
   54 GMEAN=DIF(J)+GMEAN
      GMEAN=GMEAN/MJ
      DO 55 J=1,MJ
   55 DIF(J)=DIF(J)-GMEAN
      NN=NSAM*3
      XMAX=0.0
      DO 56 I=1,NN
   56 IF(FMS(I).GT.XMAX) XMAX=FMS(I)
      NN=XMAX/10.0+1.0
      XMAX=NN*10.0
      TINT=XMAX/LB
      CALL COUNT(FMS,NSAM,3,0.0,TINT,LB,FREQ,XBAR,XVAR)
      XMIN=0.0
      XMAX=TINT*LB
      XBAR(4)=MJ*TVAR+EVAR
      XXXX=0.0
      DO 58 J=1,MJ
   58 XXXX=XXXX+DIF(J)**2
      DFA=MI-1
      DFB=MJ-1
      DFE=(MI-1)*(MJ-1)
      XXXX=XXXX/DFB
      XBAR(5)=EVAR+XXXX*MI
      XBAR(6)=EVAR
      XVAR(4)=(2.0*(MJ*TVAR+EVAR)**2)/DFA
      XVAR(5)=((EVAR+2.0*MI*XXXX    )*2.0*EVAR)/DFB
      XVAR(6)=(2.0*EVAR*EVAR)/DFE
      CALL DISCRP(3,XBAR,XVAR,LAB(1),52,52H MEAN SQUARES AND EXPECTED VA
     1LUES UNDER ANOVA MODEL )
```

```
FORTRAN IV G COMPILER        MAIN             09-15-71          15:54.53       PAGE 0004

              CALL VARXX(NSAM,MJ,BSUM,BSS)
              WRITE(6,106)
              CO 63 J=1,MJ
              XX=0.0
              DO 61 M=1,MJ
              D1=-1.0/MJ
              IF(J.EQ.M) D1=D1+1.0
              DO 60 K=1,MJ
              D2=-1.0/MJ
              IF(J.EQ.K) D2=D2+1.0
              MK=MJ*(M-1)+K
           60 XX=XX+D1*D2*DIS(MK)
           61 CONTINUE
              XX=XX/MI
           63 WRITE(6,104) J,BSUM(J),DIF(J),BSS(J),XX
              II=NSAM*2
              CO 70 I=1,NSAM
              II=II+1
           70 FMS(I)=1.0-FMS(II)/FMS(I)
              CALL RANK(FMS(1),FMS(NSAM+1),NSAM)
              IF(MODE.NE.1) CALL RELDIS(FMS,NSAM,DFA,DFE,FREQ,LB,REL,TEMP,SIG,
             1XBAR,XVAR,IPUNCH,IPLOT,0,LAB)
              IF(MODE.NE.0) CALL RELDIS(FMS,NSAM,DFA,DFE,FREQ,LB,REL,TEMP,SIG,
             1XBAR,XVAR,IPUNCH,IPLOT,1,LAB)
              WRITE(6,105) IX
              GO TO 10
           99 STOP
              END

TOTAL MEMCRY REQUIREMENTS 017A4C BYTES
```

```
FORTRAN IV G COMPILER        CATA              09-15-71          15:55.01        PAGE 0001

                SUBROUTINE DATA(MI,MJ,BS,ERR,RR,Y,X,TEMP,IX,IDIST,ISISE,XBAR,R,
               1XVAR)
           C   PURPOSE         CREATES DATA MATRIX FOR REL02
           C        MI         SAMPLE SIZE
           C        MJ         NO OF ITEMS
           C        BS         INPUT VECTOR OF BISERIAL CORRELATIONS
           C        ERR        A VECTOR OF STANDARD DEVIATION OF ERRORS
           C        RR         A VECTOR OF THRESHOLD CONSTANTS FOR EACH ITEM
           C        Y          DATA MATRIX
           C        X          WORKING VECTOR
           C        TEMP       A WORKING VECTOR
           C        IX         SEED RANDOM NUMBER
           C        IDIST      OPTION FOR TRUE SCORE DISTRIBUTION
           C        IDISE      OPTION FOR ERROR SCORE DISTRIBUTION
           C        XBAR       SUM OF SCORES FOR EACH ITEM
           C        R          INTER ITEM SUM OF PRODUCTS
           C        XVAR       A VECTOR OF PARALLEL ITEM SUM OF PRODUCTS
           C    DIMENSION R(MJ*(MJ+1)/2),XBAR(2*MJ)
                DIMENSION BS(MJ),ERR(MJ),RR(MJ),Y(MI,MJ),X(MI,MJ),TEMP(MI),R(1),
               1XBAR(1),XVAR(MJ)
                MM=MI*MJ
                IF(IDISE) 10,10,12
             10 CALL BOXSN(Y,MM,IX)
                CALL BOXSN(X,MM,IX)
                GO TO 15
             12 CALL DISE(Y,MI,MJ,IX)
                CALL DISE(X,MI,MJ,IX)
             15 IF(IDIST) 16,16,18
             16 CALL BOXSN(TEMP,MI,IX)
                GO TO 19
             18 CALL DIST(TEMP,MI,IX)
             19 CONTINUE
                DO 70 I=1,MI
                DO 20 J=1,MJ
                CUT=RR(J)
                TR=BS(J)*TEMP(I)
                YY=TR+ERR(J)*Y(I,J)
                Y(I,J)=0.0
                IF(YY.GE.CUT) Y(I,J)=1.0
                YY=TR+ERR(J)*X(I,J)
                X(I,J)=0.0
                IF(YY.GE.CUT) X(I,J)=1.0
             20 CONTINUE
                DO 30 J=1,MJ
                XBAR(J)=XBAR(J)+Y(I,J)
                JJ=MJ+J
             30 XBAR(JJ)=XBAR(JJ)+X(I,J)
                IR=0
                DO 50 J=1,MJ
                DO 40 K=1,J
                IR=IR+1
             40 R(IR)=R(IR)+Y(I,K)*Y(I,J)
             50 XVAR(J)=XVAR(J)+Y(I,J)*X(I,J)
             70 CONTINUE
                RETURN
```

```
FORTRAN IV G COMPILER        CATA          09-15-71        15:55.01        PAGE 0002
 0038                END
  TOTAL MEMCRY REQUIREMENTS 000730 BYTES
```

```
FORTRAN IV G COMPILER          EXAMPL          09-15-71          15:55.03          PAGE 0001

0001              SUBROUTINE EXAMPL(MI,MJ,BS,ERR,TEMP,RR,Y,IX,IDIST,IDISE,X,XBAR,BB,
                 1FMS,REL,R,XVAR)
            C        MI          SAMPLE SIZE
            C        MJ          NO OF ITEMS IN THE TEST
            C        BS          A VECTOR OF BISERIAL CORRELATIONS FOR EACH ITEM
            C        ERR         A VECTOR OF STANDARD DEVIATION OF ERRORS
            C        TEMP        A WORKING VECTOR
            C        RR          A VECTOR OF THRESHOLD CONSTANTS FOR EACH ITEM
            C        Y           DATA MATRIX
            C        IX          SEED RANDOM NUMBER
            C        IDIST       OPTION FOR TRUE SCORE DISTRIBUTION
            C        IDISE       OPTION FOR ERROR SCORE DISTRIBUTION
            C        X           WORKING VECTOR
            C        XVAR        SUM OF ITEM SCORE
            C        BB          WORKING VECTOR
            C        FMS         WORKING MATRIX
            C        REL         POPULATION RELIABILITY
            C        R           WORKING VECTOR
            C        XVAR        SUM OF PRODUCTS OF PARALLEL ITEMS
            C    SUBPROGRAM      DATA,MXOUT,ANOV,VEOUT,DISP,ROZB
            C         DIMENSION XBAR(2*MJ),R(MJ*(MJ+1)/2)
0002              DIMENSION BS(MJ),ERR(MJ),TEMP(MI),RR(MJ),Y(MI,MJ),BB(MJ),
                 1FMS(MJ,MJ),X(MI,MJ),XBAR(1),R(1),XVAR(MJ)
0003          100 FORMAT(1H1,20('@'),/,1X,'@',3X,'EXAMPLE RUNS',3X,'@',/,1X,20('@'))
0004          101 FORMAT(/,1X,'MSA=',E14.6,2X,'MSB=',E14.6,2X,'MSE=',E14.6,2X,'F=',
                 1E14.6,2X, 'KR20=',F8.5,2X,'UNBIASED REL EST(ANOVA)=',F8.5)
0005          102 FORMAT(/,1X,'SAMPLE DISPERSION : SATURATION COEFF=',F9.5,3X,'HOMOG
                 1ENEITY COEFF=',F9.5,5X,'HOM/SAT=',F9.5)
0006          103 FORMAT(/,1X,'GMEAN=',E14.6)
0007          104 FORMAT(/,1X,'VARIANCE OF ALPHA ESTIMATE UNDER ANOVA=',F8.5,3X,
                 1'ESTIMATE=',F8.5)
0008          105 FORMAT(/,1X,'VARIANCE OF UNBIASED REL ESTIMATES UNDER ANOVA=',F8.5
                 1,3X,'ESTIMATE=',F8.5)
0009              WRITE(6,100)
0010              CALL DATA(MI,MJ,BS,ERR,RR,Y,X,TEMP,IX,IDIST,IDISE,XBAR,R,XVAR)
0011              CALL MXOUT(Y,MI,MJ,0,12,12HDATA MATRIX )
0012              CALL ANOV(Y,MI,MJ,FMSA,FMSB,FMSE,BB)
0013              FF=FMSA/FMSE
0014              AL=1.0-1.0/FF
0015              ALL=(2.0+(MI-3.0)*AL)/(MI-1.0)
0016              WRITE(6,101) FMSA,FMSB,FMSE,FF,AL,ALL
0017              CALL VEOUT(BB,MJ,28,28HSAMPLE FIXED EFFECTS VECTOR )
0018              CALL DISP(Y,MI,MJ,FMS,BB)
0019              CALL VEOUT(BB,MJ,20,20HSAMPLE MEANS VECTOR )
0020              SUM=0.0
0021              DO 20 J=1,MJ
0022           20 SUM=SUM+BB(J)
0023              SUM=SUM/MJ
0024              WRITE(6,103) SUM
0025              CALL MXOUT(FMS,MJ,MJ,0,24,24HSAMPLE DISPERSION MATRIX)
0026              CALL ROZB(FMS,MJ,SAT,HOM)
0027              ALPHA=HOM/SAT
0028              WRITE(6,102) SAT,HOM,ALPHA
0029              IF(MI.LE.5) GO TO 90
0030              VF=(2.0*(MI-1.0)*(MJ*MI-MJ-2.0))/((MI-5.0)*(MJ-1.0)*(MI-3.0)**2)
```

```
FORTRAN IV G COMPILER        EXAMPL          09-15-71          15:55.03        PAGE 0002

                VARA=VF*(1-REL)**2
                VARE=VF*(1-AL)**2
                WRITE(6,104) VARA,VARE
                C2=(MI-3.0)/(MI-1.0)
                VARA=VARA*C2*C2
                VARE=VARE*C2*C2
                WRITE(6,105) VARA,VARE
             90 RETURN
                END

  TOTAL MEMCRY REQUIREMENTS 000A58 BYTES
```

```
FORTRAN IV G COMPILER        ITEMCO          09-15-71          15:55.06        PAGE 0001

              SUBROUTINE ITEMCO(X,Y,RXY,D1,D2,COV)
        C   PURPOSE          CALCULATE INTER-ITEM COVARIANCE FOR REL02 BASED ON
        C                    TCHEBYCHEFF-HERMITE POLINOMIALS UNDER NORMAL OGIVE
        C                    MODEL
        C        X           THRESHOLD CONST FOR FIRST ITEM
        C        Y           THRESHOLD CONST FOR SECOND ITEM
        C        RXY         INTER ITEM TETRACHORIC CORRELATION
        C        D1          NORMAL DENSITY AT Z=X
        C        D2          NORMAL DENSITY AT Z=Y
        C        COV         OUTPUT COVARIANCE
        C 200 FORMAT(1X,I3,E16.8)
              REAL*8 X1,X2,Y1,Y2,RRR,DDD,RN,U,V,X3,Y3,DD2,FI
              U=X
              V=Y
              X1=1.0
              X2=X
              Y1=1.0
              Y2=Y
              FI=2.0
              RN=DLOG(FI)
              RRR=RXY
              DDD=RRR+(U*V*RRR*RRR)/2.0
              RRR=DLOG(RRR)
              DO 10 I=3,20
              FI=I
              X3=U*X2-(FI-2.0)*X1
              X1=X2
              X2=X3
              Y3=V*Y2-(FI-2.0)*Y1
              Y1=Y2
              Y2=Y3
              RN=RN+DLOG(FI)
              DD2=X3*Y3*(DEXP(FI*RRR-RN))
              DDD=DDD+DD2
        C     COV=DDD*D1*D2
        C     WRITE(6,200) I,COV
           10 CONTINUE
              COV=DDD*D1*D2
              RETURN
              END

 TOTAL MEMORY REQUIREMENTS 000386 BYTES
```

```
FORTRAN IV G COMPILER          PARALL          09-15-71          15:55.07          PAGE 0001

 0001             SUBROUTINE PARALL(R,XBAR,MI,MJ,NSAM,XVAR,FC,FERR,COR)
            C   PURPOSE        CALCULATE TEST VARIANCE AND COVARIANCE BY PARALLEL
            C                  METHOD
            C        R         INPUT INTER ITEM SUM OF PRODUCTS
            C        XBAR      INPUT SUM OF ITEM SCORES
            C        MI        SAMPLE SIZE
            C        MJ        NO OF ITEMS IN THE TEST
            C        NSAM      NO OF SIMULATION RUNS
            C        XVAR      SUM OF PRODUCTS OF PARALLEL ITEMS
            C        FC        OUTPUT TRUE SCORE VARIANCE
            C        FERR      OUTPUT ERROR SCORE VARIANCE
            C        COR       OUTPUT RELIABILITY BETWEEN PARALLEL TESTS
            C     DIMENSION R(MJ*(MJ+1)/2),XBAR(2*MJ)
 0002             DIMENSION R(1),XBAR(1),XVAR(MJ)
 0003         100 FORMAT(1H1,60('@')/,1X,'@',2X,'ESTIMATION OF POPULATION PARAMETERS
                 1 BY PARALLEL METHOD',2X,'@',/,1X,60('@'),/,1X,'MEAN',16X,E14.6,/,
                 21X,'VARIANCE',12X,E14.6,/,1X,'TRUE VARIANCE',7X,E14.6,/,1X,'ERROR
                 3VARIANCE',6X,E14.6,/,1X,'RELIABILITY',10X,F9.6,/,1X,'KR20',17X,
                 4F9.6,/,1X,'NO OF CASES',11X,I8)
 0004             NNN=(NSAM+1)*MI
 0005             DFN=NNN-1.0
 0006             FC=0.0
 0007             FV=0.0
 0008             FD=0.0
 0009             IR=0
 0010             DO 15 J=1,MJ
 0011             DO 14 K=1,J
 0012             IR=IR+1
 0013             R(IR)=(R(IR)-(XBAR(K)*XBAR(J))/NNN)/DFN
 0014             FV=FV+R(IR)
 0015          14 CONTINUE
 0016             FD=FD+R(IR)
 0017             JJ=MJ+J
 0018             XVAR(J)=(XVAR(J)-(XBAR(J)*XBAR(JJ)/NNN))/DFN
 0019             FC=FC+XVAR(J)
 0020          15 CONTINUE
 0021             FC=2.0*(FV-FD)+FC
 0022             FV=2.0*(FV-FD)+FD
 0023             F20=(MJ*(1.0-FD/FV))/(MJ-1.0)
 0024             FMEAN=0.0
 0025             DO 20 J=1,MJ
 0026             XBAR(J)=XBAR(J)/NNN
 0027          20 FMEAN=FMEAN+XBAR(J)
 0028             COR=FC/FV
 0029             FERR=FV-FC
 0030             WRITE(6,100) FMEAN,FV,FC,FERR,COR,F20,NNN
 0031             CALL VEOUT(XBAR,MJ,12,12HMEAN VECTOR )
 0032             CALL VEOUT(XVAR,MJ,28,28HPARALLEL ITEM COVARIANCES    )
 0033             CALL MXOUT(R,MJ,MJ,1,32,32HWITHIN TEST DISPERSION MATRIX    )
 0034             RETURN
 0035             END

  TOTAL MEMCRY REQUIREMENTS CC0692 BYTES
```

```
FORTRAN IV G COMPILER        POPR              09-15-71          15:55.08          PAGE 0001

0001              SUBROUTINE POPR(MJ,BS,DIF,R,REL,ALPHA,FMT,RR,PP,ERR,TEMP,TVAR,
                 1EVAR,DIS)
             C   PURPCSE         PERFORMS BASIC COMPUTATIONS FOR REL02 POPULATICN
             C                   PARAMETERS
             C       MJ          NO OF ITEMS
             C       BS          A VECTOR OF BISERIAL CORRELATIONS
             C       DIF         A VECTOR OF ITEM DIFFICULTY
             C       R           INTER ITEM CORRELATION MATRIX
             C       REL         POPULATION RELIABILITY
             C       ALPHA       POPULATION ALPHA COEFFICIENT
             C       FMT         FORMAT FOR INPUT VECTORS
             C       RR          A VECTOR OF THRESHOLD CONSTANTS FOR EACH ITEM
             C       PP          A VECTOR OF ITEM DIFFICULTY ,REPLACED BY S.D.
             C       ERR         A VECTOR OF STANDARD DEVIATION OF ERRORS
             C       TEMP        A WORKING VECTOR
             C       TVAR        POPULATION TRUE VARIANCE UNDER N.O. MODEL
             C       EVAR        POPULATION ERROR VARIANCE UNDER N.O. MODEL
             C       DIS         OUTPUT INTER ITEM DISPERSION MATRIX
0002              DIMENSION BS(MJ),DIF(MJ),R(MJ,MJ),FMT(20),RR(MJ),PP(MJ),ERR(MJ),
                 1TEMP(MJ),DIS(MJ,MJ)
0003          100 FORMAT(/,/,1X,'@@@@@  ',I3,'TH ITEM DIFFICULTY IS LESS THAN 0.0 OR
                 1 GREATER THAN 1.0 DIF=',E14.5)
0004          101 FORMAT(/,1X,'POPULATION PARAMETERS UNDER NORMAL OGIVE MODEL')
0005          102 FORMAT(1X,'MEAN=',E14.6,3X,'VAR=',E14.6,3X,'TRUE VAR=',E14.6,3X,
                 1'ERROR VAR=',E14.6,3X,'REL=',F7.5,3X,'KR20=',F7.5)
0006          103 FORMAT(/,1X,'ITEM PARAMETERS',/,1X,'ITEM',2X,'BIS COR',10X,'DIFFIC
                 1ULTY',7X,'VARIANCE',9X,'THRES CONS.',6X,'DISC. POWER',6X,'DIFF. IN
                 2DEX')
0007          104 FORMAT(1X,I3,1X,6(E14.6,3X))
0008          105 FORMAT(20A4)
0009          106 FORMAT(/,1X,'FORMAT FOR ITEM PARAMETERS',5X,20A4)
0010              READ(5,105) (FMT(I),I=1,20)
0011              WRITE(6,106) (FMT(I),I=1,20)
0012              READ(5,FMT) (DIF(J),J=1,MJ)
0013              READ(5,FMT) (BS(J),J=1,MJ)
0014              DO 10 J=1,MJ
0015              DO 10 I=1,J
0016              R(I,J)=BS(J)*BS(I)
0017           10 R(J,I)=R(I,J)
0018              CALL MXOUT(R,MJ,MJ,0,44,44HINTER ITEM TETRACHORIC CORRELATION MATR
                 1IX    )
0019              WRITE(6,103)
0020              DO 25 J=1,MJ
0021              BIS=BS(J)
0022              DIFF=DIF(J)
0023              PP(J)=DIFF*(1.0-DIFF)
0024              ERR(J)=SQRT(1.0-BIS*BIS)
0025              AA=BIS/ERR(J)
0026              CALL NDTRI(DIFF,RRR,TEMP(J),IER)
0027              IF(IER.NE.0) WRITE(6,100) J,DIFF
0028              RRR=-RRR
0029              BB=RRR/(AA*ERR(J))
0030              RR(J)=RRR
0031           25 WRITE(6,104) J,BIS,DIFF,PP(J),RR(J),AA,BB
0032              DO 31 J=1,MJ
```

```
          DO 30 I=1,J
          RXY=BS(I)*BS(J)
          CALL ITEMCO(RR(I),RR(J),RXY,TEMP(I),TEMP(J),DIS(I,J))
          DIS(J,I)=DIS(I,J)
       30 CONTINUE
       31 CONTINUE
          SUM=0.0
          COV=0.0
          SSS=0.0
          REL=0.0
          DO 33 J=1,MJ
          DO 32 K=1,J
       32 REL=REL+DIS(K,J)
          SUM=SUM+DIF(J)
          TEMP(J)=DIS(J,J)
          COV=COV+TEMP(J)
          DIS(J,J)=PP(J)
          SSS=SSS+PP(J)
       33 CONTINUE
          TVAR=2.0*(REL-COV)+SSS
          REL=REL*2-COV
          REL=REL/TVAR
          ALPHA=(MJ*(1.0-SSS/TVAR))/(MJ-1.0)
          SSS=TVAR*REL
          EVAR=TVAR-SSS
          WRITE(6,101)
          WRITE(6,102) SUM,TVAR,SSS,EVAR,REL,ALPHA
          CALL VEOUT(TEMP,MJ,28,28HPARALLEL ITEM COVARIANCES   )
          CALL MXOUT(DIS,MJ,MJ,0,28,28HINTER ITEM DISPERSION MATRIX)
          DO 51 J=1,MJ
          PP(J)=SQRT(PP(J))
          DO 50 K=1,J
          R(J,K)=DIS(J,K)/(PP(J)*PP(K))
       50 R(K,J)=R(J,K)
       51 CONTINUE
          CALL MXOUT(R,MJ,MJ,0,32,32HINTER ITEM CORRELATION MATRIX   )
          RETURN
          END
```

TOTAL MEMCRY REQUIREMENTS 000DAE BYTES

```
FORTRAN IV G COMPILER        DIST              09-15-71          15:55.13        PAGE 0001

 0001                 SUBROUTINE DIST(TEMP,MI,IX)
            C    PURPCSE          CREATE STANDARD RANDOM TRUE SCORE MATRIX FOR REL02
            C        TEMP         OUTPUT TRUE SCORE VECTOR
            C        MI           SAMPLE SIZE, LENGTH OF TEMP
            C        IX           SEED ODD INTEGER RANDCM NUMBER
            C****THIS EXAMPLE PRODUCES EXPONENTIAL TRUE OR LATENT SCORES
 0002                 DIMENSION TEMP(MI)
 0003                 CALL VECRAN(TEMP,MI,IX)
 0004                 CO 10 I=1,MI
 0005              10 TEMP(I)=-ALCG(TEMP(I))-1.0
 0006                 RETURN
 0007                 END

  TOTAL MEMCRY REQUIREMENTS 000104 BYTES
```

```
FORTRAN IV G COMPILER        DISE              09-15-71          15:55.13        PAGE 0001

 0001                SUBROUTINE DISE(Y,MI,MJ,IX)
            C    PURPOSE         CREATE STANDARD RANDOM ERROR MATRIX Y FOR REL02
            C         Y          OUTPUT MATRIX
            C         MI         NO OF ROWS OF Y
            C         MJ         NO OF CCLS OF Y
            C         IX         SEED ODD INTEGER RANDOM NUMBER
 0002                DIMENSION Y(MI,MJ)
 0003                CALL VECRAN(Y,(MI*MJ),IX)
            C        SQR=SQRT(12.0)
 0004                SQR=SQRT(12.0)
 0005                DO 20 J=1,MJ
 0006                DO 10 I=1,MI
 0007             10 Y(I,J)=(Y(I,J)-0.5)*SQR
            C****THIS EXAMPLE PRODUCES UNIFORM ERROR SCORES
 0008             20 CONTINUE
 0009                RETURN
 0010                END

  TOTAL MEMCRY REQUIREMENTS 000264 BYTES
15:55.14 14.494 RC=0
```

SUBROUTINE PACKAGE RELOO

FORTRAN IV G COMPILER   ANOV   09-15-71   15:55.18   PAGE 0001

```
      SUBROUTINE ANOV(Y,MI,MJ,FMSA,FMSB,FMSE,BB)
C   PURPOSE         CALCULATE MEAN SQUARES AND PART-TSEST MEANS FOR RELO1
C      Y            INPUT DATA MATRIX
C      MI           SAMPLE SIZE
C      MJ           NO OF PARTS
C      FMSA         MEAN SQUARES FOR SUBJECT EFFECTS
C      FMSB         MEAN SQUARES FOR ITEM EFFECTS
C      FMSE         MEAN SQUARES FOR ERRORS
C      BB           OUTPUT ITEM MEAN VECTOR
      DIMENSION Y(MI,MJ),BB(MJ)
      S1=0.0
      S2=0.0
      S3=0.0
      S4=0.0
      DO 15 I=1,MI
      FMSA=0.0
      DO 12 J=1,MJ
   12 FMSA=FMSA+Y(I,J)
      S4=S4+FMSA
   15 S2=S2+FMSA**2
      FMSB=S4/(MI*MJ)
      S2=S2/MJ
      S4=(S4*S4)/(MI*MJ)
      DO 30 J=1,MJ
      FMSA=0.0
      DO 25 I=1,MI
      S1=S1+Y(I,J)**2
   25 FMSA=FMSA+Y(I,J)
      BB(J)=FMSA/MI-FMSB
      S3=S3+FMSA**2
   30 CONTINUE
      S3=S3/MI
      FMSA=(S2-S4)/(MI-1.0)
      FMSB=(S3-S4)/(MJ-1.0)
      FMSE=(S1-S2-S3+S4)/((MI-1.0)*(MJ-1.0))
      RETURN
      END
```

TOTAL MEMORY REQUIREMENTS 0004CA BYTES

```
FORTRAN IV G COMPILER        BOXSN           09-15-71          15:55.19        PAGE 0001

0001                SUBROUTINE BOXSN(Z,N,IX)
              C   PURPOSE          GENERATE STANDARD RANDOM NORMAL VECTOR
              C        Z           OUTPUT VECTOR OF RANDOM NUMBERS
              C        N           LENGTH OF Z
              C        IX          SEED ODD INTEGER RANDOM NUMBER
              C   SUBPROGRAMS      VECRAN
              C   METHOD           BOX-MULLER, ANN. MATH. STAT. 1959
              C     DIMENSION Z(2*NN)  NN=(N+1)/2
0002                DIMENSION Z(1)
0003                PAI=6.283185307
0004                NN=(N+1)/2
0005                CALL VECRAN(Z,(NN*2),IX)
0006                DO 20 I=1,NN
0007                XX=PAI*Z(I)
0008                II=I+NN
0009                YY=SQRT(-2.0*ALOG(Z(II)))
0010                Z(I)=YY*COS(XX)
0011                Z(II)=YY*SIN(XX)
0012             20 CONTINUE
0013                RETURN
0014                END

 TOTAL MEMORY REQUIREMENTS 00027E BYTES
```

```
FORTRAN IV G COMPILER        CHIPRB           09-15-71          15:55.20         PAGE 0001

                 FUNCTION CHIPRB(CHI,NDF)
         C   PURPOSE          CALCULATE PROBABILITY OF CHI-SQUARE VARIATE EXCEEDING
         C                    INPUT VALUE
         C       CHI          INPUT VALUE
         C       NDF          DEGREES OF FREEDOM
         C   PROGRAMMER       D. FLATHMAN
                 EXTERNAL ERF,SQRT
                 REAL NORMAL
                 INTEGER F
                 LOGICAL BIGX,EVEN
                 NORMAL(X)=0.5*(1.0+ERF(0.7071068*X))
                 F=NDF
                 X=CHI
                 CHIPRB=1.0
                 IF(X.LE.0..OR.F.LT.1) RETURN
                 A=0.5*X
                 BIGX=A.GT.10.
                 EVEN=(2*(F/2)-F).EQ.0
                 IF(EVEN.OR.(F.GT.2.AND..NOT.BIGX)) Y=EXP(-A)
                 IF(EVEN) S=Y
                 IF(.NOT.EVEN) S=2.0*NORMAL(-SQRT(X))
                 CHIPRB=S
                 IF(F.LE.2) RETURN
                 X=0.5*(F-1.0)
                 IF(EVEN) Z=1.0
                 IF(.NOT.EVEN) Z=0.5
                 IF(.NOT.BIGX) GO TO 2
                 IF(EVEN) E=0.
                 IF(.NOT.EVEN) E=0.5723649
                 C=ALOG(A)
          1      E=ALOG(Z)+E
                 S=EXP(C*Z-A-E)+S
                 Z=Z+1.0
                 IF(Z.LE.X) GO TO 1
                 CHIPRB=S
                 RETURN
          2      IF(EVEN) E=1.0
                 IF(.NOT.EVEN) E=0.5641896/SQRT(A)
                 C=0.
          3      E=E*A/Z
                 C=C+E
                 Z=Z+1.0
                 IF(Z.LE.X) GO TO 3
                 CHIPRB=C*Y+S
                 RETURN
                 END

 TOTAL MEMCRY REQUIREMENTS 00055E BYTES
```

```
FORTRAN IV G COMPILER        COUNT          09-15-71          15:55.23        PAGE 0001

              SUBROUTINE COUNT(X,LX,NV,XMIN,TINT,LB,FREQ,XBAR,XVAR)
      C     PURPOSE           CALCULATE FREQUENCY DISTRIBUTIONS
      C          X            INPUT DATA MATRIX
      C          LX           NO OF OBSERVATIONS
      C          NV           NO OF VARIABLES
      C          XMIN         INPUT MINIMUM VALUE ASSUMED
      C          TINT         INPUT CLASS INTERVAL
      C          LB           INPUT NO OF CLASS INTERVALS
      C          FREQ         OUTPUT FREQUENCY DISTRIBUTIONS
      C          XBAR         OUTPUT MEAN VECTOR
      C          XVAR         OUTPUT VARIACE VECTOR
              DIMENSION X(LX,NV),FREQ(LB,NV),XBAR(NV),XVAR(NV)
              XMAX=XMIN+TINT*LB
          100 FORMAT(/,1X,'MAXIMUM=',E14.6,3X,'MINUMUM=',E14.6,3X,'CLASS  INTERVA
             1L=',E14.6)
              WRITE(6,100) XMAX,XMIN,TINT
              DO 15 J=1,NV
              DO 10 I=1,LB
           10 FREQ(I,J)=0.0
              XBAR(J)=0.0
           15 XVAR(J)=0.0
              DO 80 J=1,NV
              DO 70 I=1,LX
              XBAR(J)=XBAR(J)+X(I,J)
              XVAR(J)=XVAR(J)+X(I,J)*X(I,J)
              XXL=XMIN
              DO 60 K=1,LB
              XXU=XXL+TINT
              IF(X(I,J).GE.XXL.AND.X(I,J).LT.XXU) GO TO 68
           60 XXL=XXU
              GO TO 70
           68 FREQ(K,J)=FREQ(K,J)+1.0
           70 CONTINUE
           80 CONTINUE
              CALL VARXX(LX,NV,XBAR,XVAR)
      C       CALL MXOUT(FREQ,LB,NV,0,24,24HFREQUENCY DISTRIBUTIONS )
              RETURN
              END

TOTAL MEMORY REQUIREMENTS 00055A BYTES
```

```
FORTRAN IV G COMPILER        CISCRP           09-15-71         15:55.24        PAGE 0001

                     SUBROUTINE DISCRP(N,XBAR,XVAR,LAB,NUMHOL,TITLE)
               C    PURPOSE           OUTPUT DISCRIPTIVE TABLE
               C        N             NO OF VARIABLES
               C        XBAR          MEAN VECTORS
               C        XVAR          VARIANCES
               C        LAB           LABELS
               C        NUMHOL        NO OF CHARACTERS IN TITLE(MULTIPLE OF 4)
               C        TITLE         TITLE OF THE TABLE
                 100 FORMAT(1H0,'DESCRIPTIVE STATISTICS FOR ',20A4)
                 101 FORMAT(1H0,20X,'MEAN',15X,'|',10X,'VARIANCE',/,1X,11X,'OBSERVED',
                    17X,'EXPECTED',5X,'|',5X,'OBSERVED',8X,'EXPECTED')
                 102 FORMAT(1X,A8, 1X,2E14.6,2X,'|',3X,2E14.6)
                     DIMENSION XBAR(N),XVAR(N),LAB(N),TITLE(20)
                     REAL*8 LAB
                     NN=(NUMHCL+3)/4
                     WRITE(6,10C) (TITLE(J),J=1,NN)
                     WRITE(6,101)
                     DO 10 I=1,N
                     II=I+N
                     WRITE(6,102) LAB(I),XBAR(I),XBAR(II),XVAR(I),XVAR(II)
                  10 CCNTINUE
                     RETURN
                     END

  TOTAL MEMCRY REQUIREMENTS 0C0368 BYTES
```

```
FORTRAN IV G COMPILER        DISP            09-15-71         15:55.26        PAGE 0001

                   SUBROUTINE DISP(Y,MI,MJ,S,XBAR)
C    PURPOSE            CALCULATE SAMPLE DISPERSION MATRIX AND MEAN VECTOR
C        Y              INPUT DATA MATRIX
C        MI             NO OF ROWS OF Y
C        MJ             NO OF COLS OF Y
C        S              OUTPUT SAMPLE DISPERSION MATRIX
C        XBAR           SAMPLE MEAN VECTOR
      DIMENSION Y(MI,MJ),S(MJ,MJ),XBAR(MJ)
      DO 15 J=1,MJ
      DO 10 K=1,MJ
   10 S(K,J)=0.0
   15 XBAR(J)=0.0
      DO 30 I=1,MI
      DO 25 J=1,MJ
      DO 20 K=1,J
   20 S(K,J)=S(K,J)+Y(I,K)*Y(I,J)
   25 XBAR(J)=XBAR(J)+Y(I,J)
   30 CONTINUE
      DO 50 J=1,MJ
      DO 45 K=1,J
      S(K,J)=(S(K,J)-(XBAR(K)*XBAR(J))/MI)/(MI-1.0)
   45 S(J,K)=S(K,J)
   50 CONTINUE
      DO 60 J=1,MJ
   60 XBAR(J)=XBAR(J)/MI
      RETURN
      END

 TOTAL MEMORY REQUIREMENTS 0C04FE BYTES
```

```
FORTRAN IV G COMPILER        FISHER           09-15-71           15:55.28         PAGE 0001

              FUNCTION FISHER(DFM,DFN,FR)
C    PURPOSE   CALCULATE PROBABILITY LEVEL WITH GIVEN D.F. AND F-RATIO
C         DFM        INPUT NUMERATOR D.F.
C         DFN        INPUT DENOMINATOR D.F.
C         FR         INPUT F-RATIO
          100 FORMAT(/,/,1HO,'ERROR IN FUNCTION FISHER:AN INPUT PARAMETER IS INV
             1ALID')
          101 FORMAT(1X,'INPUT F-RATIO IS LESS THAN 0.0  F=',E16.8)
          102 FORMAT(1X,'NUMERATOR D.F. IS LESS THAN 1.0 OR GREATER THAN 200,000
             1 DF1=',E16.8)
          103 FORMAT(1X,'DENOMINATOR D.F.IS LESS THAN 1.0 OR GREATER THAN 200,00
             1 DF2=',E16.8)
          104 FORMAT(1HO,'ERROR OUTPUT PROBABILITY IS INVALID DUE TO COMPUTATION
             1AL DIFFICULTY',/,1X, 'PROBABILITY IS SET AS TO -1.0E75')
          105 FORMAT(1HO,'ERROR IN CALULATING GAMMA FUNCTION')
              A=DFM/2.0
              B=DFN/2.0
              FB=((DFM*FR)/DFN)/(1.0+(DFM*FR)/DFN)
              CALL BDTR(FB,A,B,PRO,D,IER)
              IF(IER.EQ.-2) WRITE(6,100)
              IF(FR.LT.0.0) WRITE(6,101) FR
              IF(DFM.LT.1.0.OR.DFM.GT.200000) WRITE(6,102) DFM
              IF(DFN.LT.1.0.OR.DFN.GT.200000) WRITE(6,103) DFN
              IF(IER.EQ.2) WRITE(6,104)
              IF(IER.EQ.-1.OR.IER.EQ.1) WRITE(6,105)
              FISHER=1.0-PRO
              RETURN
              END

 TOTAL MEMORY REQUIREMENTS 0004F0 BYTES
```

FORTRAN IV G COMPILER FITTES 09-15-71 15:55.30 PAGE 0001

```
            SUBROUTINE FITTES(X,N,NT)
C   PURPOSE          PERFORMS CHI-SQUARE GOODNESS OF FIT TEST
C        X           INPUT FREQUNCY MATRIX
C                    COL-1 EXPECTED FREQUENCIES
C                    COL-2 OBSERVED FREQUENCIES
C        N           NO OF CLASS INTERVALS OR NO OF THE ROW OF X
C        NSAM        SAMPLE SIZE
C   SUBPROGRAM       CHIPRB
      DIMENSION X(N,2)
  100 FORMAT(/,1X,'CHI-SQ GOODNESS OF FIT TEST:',3X,'CHI=',E14.6,3X,'NDF
     1=',I5,3X,'PROB=',F8.4)
      CHI=0.0
      NDF=0
      XT=0.0
      YT=0.0
      I=1
   12 XX=0.0
      YY=0.0
   15 YY=YY+X(I,2)
      XX=XX+X(I,1)
      IF(XX.GE.5.0) GO TO 16
      I=I+1
C     IF(I.LE.N) GO TO 15
      GO TO 15
   16 NDF=NDF+1
      CHI=CHI+((XX-YY)**2)/XX
      XT=XT+XX
      YT=YT+YY
      XR=NT-XT
      I=I+1
      YR=NT-YT
C     IF(XR.GE.10.0.AND.I.LE.N) GO TO 12
      IF(XR.GE.10.0) GO TO 12
      IF(XR.GT.0.0) GO TO 20
      NDF=NDF-1
      GO TO 25
   20 CHI=CHI+((XR-YR)**2)/XR
   25 PRO=CHIPRB(CHI,NDF)
      WRITE(6,100) CHI,NDF,PRO
      RETURN
      END
```

TOTAL MEMORY REQUIREMENTS 0003CC BYTES

```
FORTRAN IV G COMPILER        FST              09-15-71          15:55.31       PAGE 0001

 0001               FUNCTION FST(DF1,DF2,P,PRE)
             C      PURPOSE      CALCULATE  F STATISTICS WHEN DEGREES OF FREEDOM AND
             C                   PROBABILITY ARE GIVEN
             C         DF1       DEGREES OF FREEDOM FOR NUMERATOR
             C         DF2       DEGREES OF FREEDOM FOR DENOMINATOR
             C         P         PROBABILITY LEVEL
             C         PRE       PRECISION LEVEL FOR OUTPUT F RATIO
             C      SUBPROGRAM   FISHER
 0002               IF (DF1.LE.0.0.OR.DF2.LE.0.0.OR.P.LE.0.0) GO TO 999
 0003           100 FORMAT (1H0,'DEGREES OF FREEDOM  OR PROBABILITY IS LESS OR EQUAL
                  1TO ZERO  RETURNS TO MAIN WITH FST=0.0')
 0004               X1=1.0
 0005               X2=0.0
 0006            10 F=(X1+X2)/2.0
 0007               FR=DF2*((1.0-F)/(DF1*F))
 0008               PRO=FISHER(DF1,DF2,FR)
 0009               ER=P-PRO
             C  101 FORMAT(1X,4F12.6)
             C      WRITE(6,101) F,FR,PRO,ER
 0010               IF(ABS(ER).LE.PRE)   GO TO 99
 0011               IF (P.LT.PRO) X1=F
 0012               IF(P.GT.PRO) X2=F
 0013               GO TO 10
 0014            99 FST=FR
 0015               RETURN
 0016           999 WRITE(6,100)
 0017               FST=0.0
 0018               RETURN
 0019               END

  TOTAL MEMCRY REQUIREMENTS 000330 BYTES
```

```
      SUBROUTINE MXOUT(A,N,M,MS,NUMHCL,TITLE)
C   PURPOSE         OUTPUTS A MATRIX
C       A           INPUT MATRIX
C       N           NO OF ROWS IN A
C       M           NO OF COLS IN A
C       MS          OPTION FOR STORAGE MODE
C                   0 GENERAL
C                   1 SYMETRIC
C                   2 DIAGONAL
C       NUMHOL      NO OF CHARACTERS OF TITLE(MULTIPLE OF 4)
C       TITLE       TITLE OF THE VECTOR IN A FORMAT
      DIMENSION A(1),B(8),TITLE(20)
  100 FORMAT(1H0,20A4)
  101 FORMAT(/,5X,8(5X,A2,I3,6X))
  102 FORMAT(1X,'R-',I3,E15.6,7E16.6)
      DATA COL/'C-'/
      NN=(NUMHOL+3)/4
      WRITE(6,100) (TITLE(J),J=1,NN)
      LINS=N+2
      J=1
      LEND=N
      NEND=8
   10 LSTRT=1
   20 CONTINUE
      JNT=J+NEND-1
      IF(JNT.GT.M) JNT=M
      WRITE(6,101)((COL,JCUR),JCUR=J,JNT)
      LTEND=LSTRT+LEND-1
      DO 80 L=LSTRT,LTEND
      DO 55 K=1,NEND
      KK=K
      JT=J+K-1
      IF(MS-1) 41,42,45
   41 IRX=N*(JT-1)+L
      GO TO 47
   42 IF(L-JT)43,44,44
   43 IRX=L+(JT*JT-JT)/2
      GO TO 47
   44 IRX=JT+(L*L-L)/2
      GO TO 47
   45 IRX=0
      IF(L-JT) 47,46,47
   46 IRX=L
   47 IJNT=IRX
      B(K)=0.0
      IF(IJNT) 50,50,49
   49 B(K)=A(IJNT)
   50 CONTINUE
      IF(JT-M) 55,60,60
   55 CONTINUE
   60 WRITE(6,102) L,(B(JW),JW=1,KK)
      IF(N-L) 85,85,80
   80 CONTINUE
      LSTRT=LSTRT+LEND
      GO TO 20
```

```
0046          85 IF(JT-M) 9C,95,95
0047          90 J=JT+1
0048             GO TO 10
0049          95 RETURN
0050             END
```

TOTAL MEMCRY REQUIREMENTS 0005A0 BYTES

```
FORTRAN IV G COMPILER        PLOT            09-15-71        15:55.37       PAGE 0001

              SUBROUTINE PLOT(X,A,N,NSAM)
      C   PURPOSE           GIVES PLOT FOR REL01
      C       X             INPUT MATRIX OF RELATIVE FREQUENCIES(%)
      C                     COL-1 EXPECTED
      C                     COL-2 OBSERVED
      C       A             WORKING VECTOR
      C       N             NO OF CLASSES(LIMITED TO 24,36 OR 48)
      C       NSAM          SAMPLE SIZE
              DIMENSION X(N,2),A(N),SIG(4)
              DATA SIG/'.','*','+',' '/
          100 FORMAT(1H1,34('%'),/,1X,'% PLOT OF FREQUENCY DISTRIBUTION %',/,1X,
             134('%'))
          101 FORMAT(4X,'FREQUENCY(%)')
          102 FORMAT(1X,F7.3,'-|',24('  ',A1,'  '))
          103 FORMAT(9X,'|',24('  ',A1,'  '))
          104 FORMAT(3X,'REL',3X,'|',24('____|'))
          105 FORMAT(4X,12(F7.1,3X),F7.1)
          106 FORMAT(/,1X,'NOTE:',3X,'(.) EXPECTED FREQ; (*) OBSERVED FREQ; (+)
             1OVERLAPPING POINT')
          107 FORMAT(1X,F7.3,'-|',36(' ',A1,' '))
          108 FORMAT(9X,'|',36(' ',A1,' '))
          109 FORMAT(3X,'REL',3X,'|',36('--|'))
          110 FORMAT(4X,13(F7.1,2X))
          111 FORMAT(1X,F7.3,'-|',48(A1,' '))
          112 FORMAT(9X,'|',48(A1,' '))
          113 FORMAT(3X,'REL',3X,'|',48('-|'))
          114 FORMAT(4X,12(F7.1,1X),F7.1)
              NNN=N-36
              WRITE(6,100)
              CALL FITTES(X,N,NSAM)
              WRITE(6,101)
              XMAX=0.0
              DO 10 I=1,N
              DO 10 J=1,2
              X(I,J)=(X(I,J)*100.0)/NSAM
              IF(X(I,J).GT.XMAX) XMAX=X(I,J)
           10 CONTINUE
              NN=XMAX/10.0+1.0
              XU=NN*10.0
              IF(XU.GT.100.0) XU=100.0
              DELT=XU/50.0
              DELTH=DELT/2.0
              XU=XU+DELTH
              DO 50 I=1,10
              XM=XU-DELTH
              DO 40 J=1,5
              XL=XU-DELT
              DO 20 K=1,N
              A(K)=SIG(4)
              IF(X( K,1).GE.XU.OR.X( K,1).LT.XL) GO TO 12
              A(K)=SIG(1)
           12 IF(X( K,2).GE.XU.OR.X( K,2).LT.XL) GO TO 20
              IF(A(K).EQ.SIG(1))GO TO 14
              A(K)=SIG(2)
              GO TO 20
```

```
FORTRAN IV G COMPILER        PLOT         09-15-71         15:55.37        PAGE 0002
      14 A(K)=SIG(3)
      20 CONTINUE
         XU=XL
         IF(NNN) 25,30,35
      25 IF(J.EQ.1) WRITE(6,102) XM,(A(K),K=1,N)
         IF(J.GT.1) WRITE(6,103) (A(K),K=1,N)
         GO TO 40
      30 IF(J.EQ.1) WRITE(6,107) XM,(A(K),K=1,N)
         IF(J.GT.1) WRITE(6,108) (A(K),K=1,N)
         GO TO 40
      35 IF(J.EQ.1) WRITE(6,111) XM,(A(K),K=1,N)
         IF(J.GT.1) WRITE(6,112) (A(K),K=1,N)
      40 CONTINUE
      50 CONTINUE
         A(1)=-0.2
         DELT=0.1
         CO 51 I=2,13
      51 A(I)=A(I-1)+DELT
         IF(NNN) 52,54,56
      52 WRITE(6,104)
         WRITE(6,105)(A(I),I=1,13)
         GO TO 58
      54 WRITE(6,109)
         WRITE(6,110) (A(I),I=1,13)
         GO TO 58
      56 WRITE(6,113)
         WRITE(6,114)(A(I),I=1,13)
      58 CONTINUE
         WRITE(6,106)
         RETURN
         END

  TOTAL MEMCRY REQUIREMENTS 000874 BYTES
```

```
FORTRAN IV G COMPILER        PUNCH            09-15-71          15:55.42        PAGE 0001

 0001               SUBROUTINE PUNCH(FREQ,LB,NV)
            C    PURPOSE        GIVES CARD OUTPUT FOR REL01
            C          LB       NO OF ROWS OF FREQ
            C          NV       NO OF COLS OF FREQ
 0002               DIMENSION FREQ(LB,NV)
 0003           100 FORMAT(I2,3X,5F10.4)
 0004               DO 20 I=1,LB
 0005            20 WRITE(7,100) I,(FREQ(I,J),J=1,NV)
 0006               RETURN
 0007               END

  TOTAL MEMORY REQUIREMENTS 000200 BYTES
```

```
      SUBROUTINE RELDIS(FMS,NSAM,DFA,DFE,FREQ,LB,REL,TEMP,SIG,XBAR,XVAR,
     1IPUNCH,IPLOT,ID,LAB)
      DIMENSION FMS(NSAM,2),FREQ(LB,2),TEMP( 130),XBAR(2),XVAR(2),LAB(4)
C   PURPOSE         INVESTIGATE SAMPLING DISTRIBUTION OF RELIABILITY
C                   ESTIMATES
C        FMS        AN INPUT MATRX
C                   COL-1 RELIABILITY ESTIMATES
C                   COL-2 RANK OF ABOVE
C        NSAM       SAMPLE SIZE
C        DFA        DEGREES OF FREEDOM OF NUMERRATOR
C        DFE        DEGREES OF FREEDOM FOR ERRORS
C        FREQ       FREQUENCY TABLE
C        LB         NO OF CLASS INTERVALS, NO OF RCWS OF FREQ
C        REL        POPULATION RELIABILITY
C        TEMP       WORKING VECTOR
C        SIG        SIGNIFICANCE LEVEL  FOR EACH TAIL
C        XBAR,XVAR  WORKING VECTORS
C        IPUNCH     OPTION FOR CARD OUTPUT
C        IPLOT      OPTION FOR PLOT
C        ID         OPTION FOR ESTIMATION FORMULA
C                   0-BIASED ALPHA FORMULA
C                   1-KRISTOF CORRECTION, UNBIASED
C        LAB        LABELS
      REAL*8 LAB
  100 FORMAT(1H1,26('a'),/,1X,'a',2X,'RELIABILITY STUDY',5X,'a',/,1X,
     126('a'))
  101 FORMAT(/,1X,'EXPECTED FREQUENCY OF RELIABILITY ESTIMATES BELOW -0.
     12=',E14.6)
  102 FORMAT(/,1X,'ESTIMATION IS BASED ON ALPHA FORMULA(ANOVA,BIASED)')
  103 FORMAT(/,1X,'ESTIMATION IS BASED ON KRISTOF CORRECTION(ANOVA,UNBIA
     1SED)')
      WRITE(6,100)
      IF(ID.EQ.0) GO TO 22
      WRITE(6,103)
      C1=2.0/DFA
      C2=(DFA-2.0)/DFA
      DO 21 I=1,NSAM
   21 FMS(I,1)=C1+C2*FMS(I,1)
      GO TO 23
   22 WRITE(6,102)
      C1=0.0
      C2=1.0
   23 CONTINUE
      FF=(1.0-REL)/1.2
      PL=FISHER(DFA,DFE,FF)
      FF=NSAM*(1.0-PL)
      WRITE(6,101) FF
      RR=-0.2
      DELT=1.2/LB
      DO 25 I=2,LB
      RR=RR+DELT
      FF=((1.0-REL)*C2)/(1.0-RR)
      PU=FISHER(DFA,DFE,FF)
      FREQ((I-1),1)=NSAM*(PL-PU)
   25 PL=PU
```

```
FORTRAN IV G COMPILER        RELDIS          09-15-71          15:55.42        PAGE 0002

                 FREQ(LB,1)=NSAM*PL
                 CALL CCUNT(FMS(1,1),NSAM,1,-0.2,DELT,LB,FREQ(1,2),XBAR,XVAR)
                 XBAR(2)=REL
                 IF(ID.EQ.0) XBAR(2)=-2.0/(DFA-2.0)+(DFA*REL)/(DFA-2.0)
                 XVAR(2)=((1.0-REL)**2)*2.0*(DFA**2)*(DFE+DFA-2.0)
                 XVAR(2)=(C2*C2*XVAR(2))/(DFE*(DFA-4.0)*(DFA-2.0)**2)
                 CALL DISCRP(1,XBAR,XVAR,LAB(4),24,24H RELIABILITY ESTIMATES  )
                 CALL MXOUT(FREQ,LB,2,0,28,28HCCMPARISON OF RELIABILITY    )
                 IF(IPUNCF.EQ.1) CALL PUNCH(FREQ,LB,2)
                 CALL SIGTES(FMS,NSAM,DFA,DFE,SIG,REL,ID)
                 IF(IPLOT.EQ.1) CALL PLOT(FREQ,TEMP,LB,NSAM)
                 RETURN
                 END

  TOTAL MEMCRY REQUIREMENTS 0C0898 BYTES
```

```
FORTRAN IV G COMPILER        ROZB             09-15-71          15:55.46        PAGE 0001

 0001                SUBROUTINE ROZB(DIS,MJ,SAT,HOM)
            C    PURPOSE          CALCULATE SATURATION AND HOMOGENEITY COEFFICIETS
            C       DIS           INPUT DISPERSION MATRIX
            C       MJ            SIZE OF DIS
            C       SAT           OUTPUT SATURATION COEFFICIENT
            C       HOM           OUTPUT HOMOGENEITY COEFFICIENT
 0002                DIMENSION DIS(MJ,MJ)
 0003                TEMP=0.0
 0004                HOM=0.0
 0005                DO 20 J=1,MJ
 0006                DO 10 K=1,MJ
 0007             10 TEMP=TEMP+DIS(K,J)
 0008             20 HOM=HOM+DIS(J,J)
 0009                SAT=TEMP/(HOM*MJ)
 0010                HOM=(TEMP-HOM)/(HOM*(MJ-1.0))
 0011                RETURN
 0012                END

  TOTAL MEMORY REQUIREMENTS 000288 BYTES
```

```
FORTRAN IV G COMPILER        SIGTES            09-15-71          15:55.47        PAGE 0001

0001              SUBROUTINE SIGTES(FMS,NSAM,DFA,DFE,SIG,REL,ID)
            C   PURPOSE          OBTAIN EMPRICAL CRITICAL POINTS OF RELIABILIRY
            C.                   ESTIMATES
            C        FMS         AN INPUT MATRX
            C                    COL-1 RELIABILITY ESTIMATES
            C                    COL-2 RANK OF ABOVE
            C        NSAM        SAMPLE SIZE
            C        DFA         DEGREES OF FREEDOM OF NUMERRATOR
            C        DFE         DEGREES OF FREEDOM FOR ERRORS
            C        SIG         SIGNIFICANCE LEVEL  FOR EACH TAIL
            C        ID          OPTION FOR ESTIMATION FORMULA
            C                    0-ALPHA COEFFICIENT
            C                    1-KRISTOF CORRECTION FOR BIASEDNESS
0002              DIMENSION FMS(NSAM,2)
0003          101 FORMAT(1X,'ALPHA ESTIMATES',3X,'SIGLEVEL(EACH)=',F5.3,3X,'DFA=',
                 1F6.0,3X,'DFE=',F8.0,16X,'LOWER BOUND=',F9.6,4X,'UPPER BOUND=',
                 2F9.6)
0004          102 FORMAT(1X,'NO OF CASES LESS THAN LOWER B=',I4,2X,F6.2,'% ; GREATER
                 1 THAN UPPER B=',I4,2X,F6.2,'%',2X,'LOWER B(EST)=',F9.6,3X,'UPPER B
                 1(EST)=',F9.6)
0005          103 FORMAT(1X,'ADJUSTEC ALPHA ESTIMATES',3X,'SIG LEVEL(EACH)=',F5.3,
                 13X,'DFA=',F6.0,3X,'DFE=',F8.0, 6X,'LOWER BOUND=',F9.6,4X,'UPPER BO
                 2CUND=',F9.6)
0006              C2=(DFA-2.0)/DFA
0007              IF(ID.EQ.0) C2=1.0
0008              FU=FST(DFE,DFA,SIG,0.0001)
0009              FL=FST(CFA,DFE,SIG,0.0001)
0010              FL=1.0/FL
0011              BL=1.0-FU*(1.0-REL)*C2
0012              BU=1.0-FL*(1.0-REL)*C2
0013              IF(ID.EQ.0) WRITE(6,101) SIG,DFA,DFE,BL,BU
0014              IF(ID.EQ.1) WRITE(6,103) SIG,DFA,DFE,BL,BU
0015              ML=0
0016              MU=0
0017              DO 10 I=1,NSAM
0018              IF(FMS(I,1).GT.BU) MU=MU+1
0019           10 IF(FMS(I,1).LT.BL) ML=ML+1
0020              EML=(ML*100.0)/NSAM
0021              EMU=(MU*100.0)/NSAM
0022              NL=NSAM*SIG+1.50001
0023              NU=NSAM*(1.0-SIG)+0.50001
0024              NLL=NL-1
0025              NUU=NU+1
0026              FNL=0.0
0027              FNLL=0.0
0028              FNU=0.0
0029              FNUU=0.0
0030              DO 20 I=1,NSAM
0031              NID=FMS(I,2)+0.500001
0032              IF(NID.EC.NL) FNL=FMS(I,1)
0033              IF(NID.EC.NLL) FNLL=FMS(I,1)
0034              IF(NID.EC.NUU) FNUU=FMS(I,1)
0035           20 IF(NID.EC.NU) FNU=FMS(I,1)
0036              SU=(FNU+FNUU)/2.0
0037              SL=(FNL+FNLL)/2.0
```

```
FORTRAN IV G COMPILER        SIGTES           09-15-71          15:55.47          PAGE 0002

 0038              WRITE(6,102) ML,EML,MU,EMU,SL,SU
 0039              RETURN
 0040              END

  TOTAL MEMORY REQUIREMENTS 000838 BYTES
```

```
FORTRAN IV G COMPILER        VARXX            09-15-71         15:55.49        PAGE 0001

 0001                SUBROUTINE VARXX(N,NV,XBAR,XVAR)
              C    PURPOSE         CALUCULATE MEANS AND VARIANCE VECTORS
              C        N           SAMPLE SIZE
              C        NV          NO OF VARIABLES
              C        XBAR        INPUT SUM OF VARIABLES, REPLACED BY MEANS
              C        XVAR        INPUT SUM OF SQUARES, REPLACED BY VARIANCES
 0002                DIMENSICN XBAR(NV),XVAR(NV)
 0003                CO 10 J=1,NV
 0004                XVAR(J)=(XVAR(J)-(XBAR(J)*XBAR(J))/N)/(N-1.0)
 0005             10 XBAR(J)=XBAR(J)/N
 0006                RETURN
 0007                END

  TOTAL MEMCRY REQUIREMENTS 000222 BYTES
```

```
FORTRAN IV G COMPILER        VECRAN          09-15-71          15:55.50        PAGE 0001

 0001              SUBROUTINE VECRAN(Z,N,IX)
            C   PURPOSE          COMPUTES N UNIFORM RANDOM NUMBERS BETWEEN 0.0 AND 1.0
            C                    USING SSP RANDU METHOD
            C        Z           OUTPUT RANDOM VECTOR
            C        N           LENGTH OF Z
            C        IX          SEED ODD INTEGER RANDOM NUMBER
            C   SUBPROGRAM       NONE
 0002              DIMENSION Z(N)
 0003              DO 20 M=1,N
 0004              IX=IX*65539
 0005              IF(IX) 5,6,6
 0006            5 IX=IX+2147483647+1
 0007            6 Y=IX
 0008              Y=Y*.4656613E-9
 0009           20 Z(M)=Y
 0010              RETURN
 0011              END

  TOTAL MEMCRY REQUIREMENTS 0001FE BYTES
```

```
FORTRAN IV G COMPILER        VEOUT            09-15-71         15:55.50        PAGE 0001

 0001                 SUBROUTINE VEOUT(A,N,NUMH,TITLE)
              C     PURPOSE          PRINTS UP A VECTOR
              C         A            INPUT VECTOR
              C         N            LENGTH OF A
              C         NUMH         NO OF CHARACTERS IN TITLE(MULTIPLE OF 4)
              C         TITLE        TITLE OF THE VECTOR
 0002                 DIMENSION A(N),TITLE(20)
 0003             100 FORMAT(/,1X,20A4)
 0004             101 FORMAT(1X,10(5X,I2,6X))
 0005             102 FORMAT(1X,10E13.5)
 0006                 NN=(NUMH+3)/4
 0007                 WRITE(6,100) (TITLE(I),I=1,NN)
 0008                 M=N
 0009                 IF(N.GT.10) M=10
 0010                 WRITE(6,101) (I,I=1,M)
 0011                 WRITE(6,102) (A(I),I=1,M)
 0012                 IF(N.LE.10) GO TO 30
 0013                 WRITE(6,101) (I,I=11,N)
 0014                 WRITE(6,102) (A(I),I=11,N)
 0015              30 RETURN
 0016                 END

  TOTAL MEMCRY REQUIREMENTS 000354 BYTES
15:55.52 17.673 RC=0
```

APPENDIX A.2

EXAMPLE OUTPUTS

REL01	:	Votaw-Jöreskog Example Data
REL02	:	Load-Novick Item Parameters

```
VOTAW-JORESKOG EXAMPLE DATA, CONGENERIC,N=2000,I=10,J=4,NORMAL
          NO OF SAMPLES SIMULATED                 2000
          NO OF SUBJECTS IN EACH SAMPLE             10
          NO OF PART-TESTS                           4
          NO OF FACTORS IN TRUE SCORE                1
          STARTING INTEGER RANDOM NUMBER          9999
          OPTION FOR CARD OUTPUT                     0
          OPTION FOR PLOT                            1
          OPTION FOR ESTIMATION FORMULA              2
          OPTION FOR THE NO OF CLASS INTERVALS      48
          SIGNIFICANCE LEVEL                     0.050
          TRUE SCORE DISTRIBUTIONS ARE NORMAL
          ERROR SCORE DISTRIBUTIONS ARE NORMAL
```

```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@   INPUT POPULATION PARAMETERS      @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
FORMAT FOR THE DATA      (4F5.4)
MEANS VECTOR
      1              2              3              4
  0.0            0.0            0.0            0.0
ERROR STANDARD DEVIATIONS VECTOR
      1              2              3              4
  0.20459E 01  0.45847E 01  0.39644E 01  0.11618E 01

FACTOR LOADING MATRIX
          C-  1
R-  1   0.457C00E 01
R-  2   0.268C00E 01
R-  3   0.265000E 01
R-  4   0.453000E 01

TRUE SCORE DISPERSION MATRIX
          C-  1            C-  2            C-  3            C-  4
R-  1   0.208849E 02    0.122476E 02    0.121105E 02    0.207021E 02
R-  2   0.122476E 02    0.718240E 01    0.710200E 01    0.121404E 02
R-  3   0.121105E 02    0.710200E 01    0.702250E 01    0.120045E 02
R-  4   0.207021E 02    0.121404E 02    0.120045E 02    0.205209E 02
TRUE SCORE DISPERSION: SATURATION COEFF=  0.93608     HOMOGENEITY COEFF=  0.91478

DISPERSION MATRIX
          C-  1            C-  2            C-  3            C-  4
R-  1   0.25C7C6E 02    0.122476E 02    0.121105E 02    0.207021E 02
R-  2   0.122476E 02    0.282019E 02    0.710200E 01    0.121404E 02
R-  3   0.121105E 02    0.710200E 01    0.227390E 02    0.120045E 02
R-  4   0.207C21E 02    0.121404E 02    0.120045E 02    0.218707E 02
PART SCORE DISPERSION: SATURATION COEFF=  0.63979     HOMOGENEITY COEF=  0.51972
POPULATION PARAMETERS
RELIABILITY                          0.83125
ALPHA                                0.81233
TRUE SCORE VARIANCE                  0.208225E 03
ERROR SCORE VARIANCE                 0.422714E 02
GMEAN=  0.0
FIXED EFFECTS
      1              2              3              4
  0.0            0.0            0.0            0.0
VARIANCES OF PARTS
      1              2              3              4
  0.25071E 02  0.28202E 02  0.22739E 02  0.21871E 02
```

```
aaaaaaaaaaaaaaaaaaaaa
a   EXAMPLE RUNS    a
aaaaaaaaaaaaaaaaaaaaa

DATA MATRIX
          C-  1            C-  2           C-  3           C-  4
R-  1   0.898699E 01   -0.589026E 01    0.240342E 01    0.315894E 01
R-  2  -0.659240E 00   -0.541119E 01   -0.618265E 01   -0.325838E 01
R-  3  -0.187743E 00   -0.114643E 01   -0.463635E 01   -0.318824E 01
R-  4   0.815396E 00   -0.833126E 01   -0.412630E 00    0.122313E 01
R-  5   0.343811E 01   -0.889207E 00    0.358485E 00    0.575484E 01
R-  6   0.775583E 01    0.440201E 01    0.340443E 00    0.247397E 01
R-  7   0.227846E 01    0.340946E 01    0.179335E 01    0.104001E 01
R-  8   0.307808E 00    0.648926E 00   -0.122800E 01   -0.585746E 01
R-  9  -0.250911E 01    0.181339E 01   -0.239520E 01   -0.152782E 01
R- 10  -0.413363E 01   -0.375924E 01    0.942052E 00   -0.433714E 01
 MSA=  0.274080E 02  MSB=  0.183557E 02  MSE=   0.986890E 01  F=   0.277721E 01  ALPHA= 0.63993  UNBIASED REL EST(ANOVA)= 0.71994
 SAMPLE FIXED EFFECTS VECTOR
       1              2              3              4
   0.19242E 01 -0.12005E 01 -0.58680E 00 -0.13691E 00
 SAMPLE MEANS VECTOR
       1              2              3              4
   0.16093E 01 -0.15154E 01 -0.90171E 00 -0.45182E 00
 GMEAN= -0.314904E 00

 SAMPLE DISPERSION MATRIX
          C-  1            C-  2           C-  3           C-  4
 R-  1   0.174050E 02    0.205353E 01    0.560189E 01    0.111896E 02
 R-  2   0.205353E 01    0.179301E 02    0.144443E 01    0.521637E 00
 R-  3   0.560189E 01    0.144443E 01    0.768107E 01    0.549751E 01
 R-  4   0.111896E 02    0.521637E 00    0.549751E 01    0.139989E 02
 SAMPLE DISPERSION : SATURATION COEFF=  0.48072   HOMOGENEITY COEFF=  0.30762     HOM/SAT=  0.63992
 VARIANCE OF ALPHA ESTIMATE UNDER ANOVA= 0.02371    ESTIMATE= 0.10796
 VARIANCE OF UNBIASED REL ESTIMATES UNDER ANOVA= 0.01434   ESTIMATE= 0.06531
```

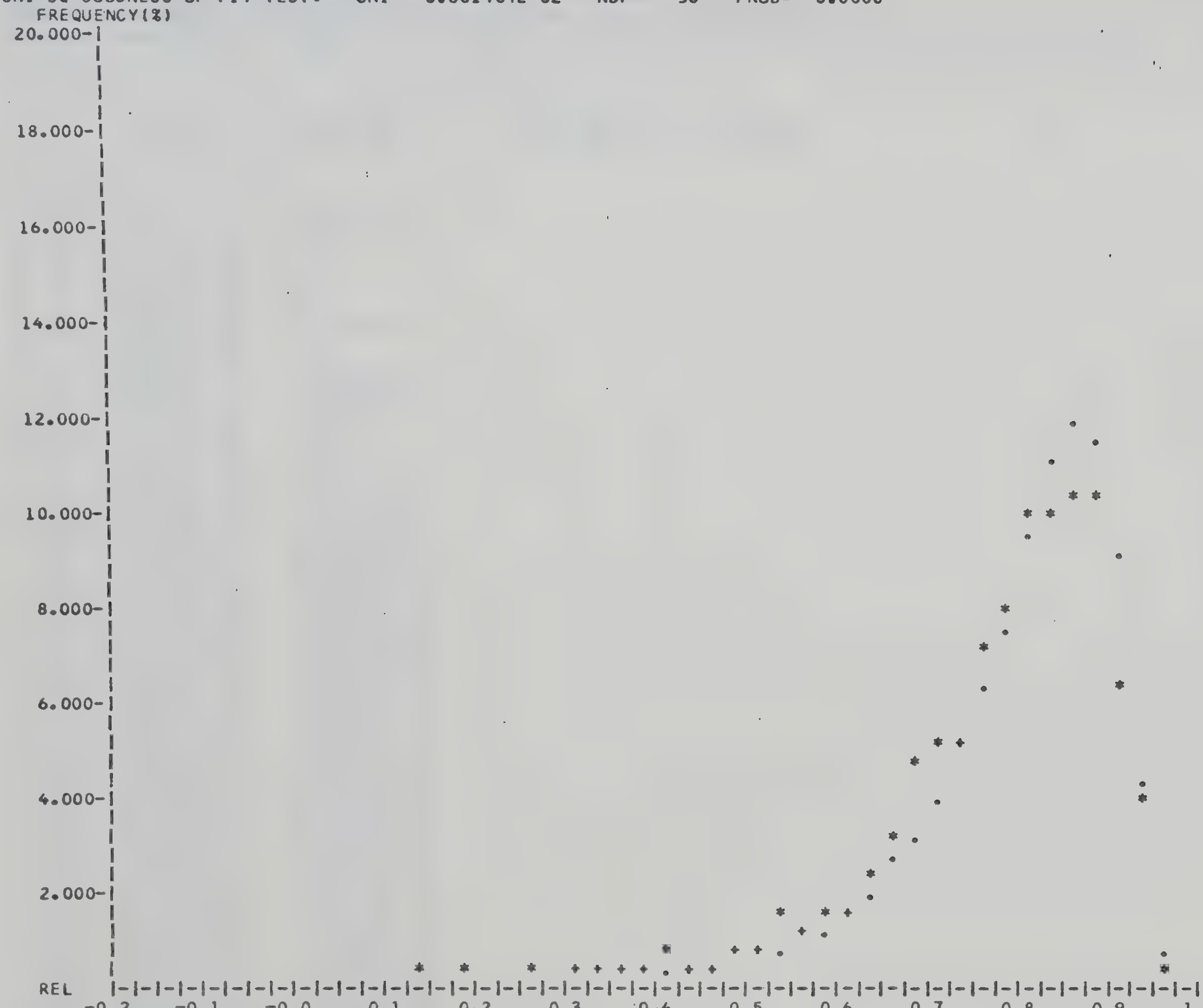
```
@@@@@@@@@@@@@@@@@@@@@@@@
@  SUMMARY OF OUTPUT  @
@@@@@@@@@@@@@@@@@@@@@@@@
MAXIMUM=  0.310000E 03   MINUMUM=  0.0           CLASS INTERVAL=  0.645833E 01

DESCRIPTIVE STATISTICS FOR  MEAN SQUARES AND EXPECTED VALUES UNDER ANOVA MODEL

                        MEAN               |            VARIANCE
            OBSERVED       EXPECTED        |     OBSERVED        EXPECTED
SUBJECT     0.620664E 02  0.626240E 02     |     0.867314E 03  0.871504E 03
PARTTEST    0.118725E 02  0.105679E 02     |     0.121838E 03  0.744530E 02
ERROR       0.116081E 02  0.105679E 02     |     0.136745E 02  0.827255E 01
DISCRIPTIVE STATISTICS FOR FIXED EFFECT ESTIMATES AND EXPECTED VALUES UNDER M.F.  MODEL
PART        MEAN          EXPECTED         |     VARIANCE        EXPECTED
   1         0.140543E-01  0.0             |     0.572055E 00  0.566122E 00
   2         0.225583E-C1  0.0             |     0.144065E 01  0.140119E 01
   3        -0.359699E-01  0.0             |     0.112026E 01  0.114170E 01
   4        -0.646006E-C3  0.0             |     0.428632E 00  0.416785E 00
```

```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@  RELIABILITY STUDY          @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
ESTIMATION IS BASED ON ALPHA FORMULA(ANOVA,BIASED)
EXPECTED FREQUENCY OF RELIABILITY ESTIMATES BELOW -0.2=  0.425398E 01
MAXIMUM=  0.999999E 00   MINUMUM= -0.2C0000E 00   CLASS INTERVAL=  0.250000E-01

DESCRIPTIVE STATISTICS FOR  RELIABILITY ESTIMATES

                       MEAN                |               VARIANCE
            OBSERVED        EXPECTED       |        OBSERVED        EXPECTED
REL COF     0.760583E 00   0.783034E 00    |        0.319032E-01    0.237113E-01

COMPARISON OF RELIABILITY
           C-  1            C-  2
R-  1    0.358701E 00     0.0
R-  2    0.396252E 00     0.0
R-  3    0.438452E 00     0.0
R-  4    0.486255E 00     0.100000E 01
R-  5    0.540137E 00     0.100000E 01
R-  6    0.601053E 00     0.100000E 01
R-  7    0.670552E 00     0.100000E 01
R-  8    0.749469E 00     0.200000E 01
R-  9    0.839591E 00     0.0
R- 10    0.9427C7E 00     0.100000E 01
R- 11    0.106120E 01     0.100000E 01
R- 12    0.119734E 01     0.200000E C1
R- 13    0.135469E 01     0.0
R- 14    0.153673E 01     0.500000E 01
R- 15    0.174820E 01     0.300000E 01
R- 16    0.199473E 01     0.400000E 01
R- 17    0.228298E 01     0.100000E 01
R- 18    0.262129E 01     0.300000E C1
R- 19    0.301957E 01     0.800000E 01
R- 20    0.349116E 01     0.300000E 01
R- 21    0.4C5061E 01     0.5000C0E C1
R- 22    0.471771E 01     0.600000E 01
R- 23    0.551689E 01     0.600000E 01
R- 24    0.647807E 01     0.600000E 01
R- 25    C.764000E 01     0.140CC0E C2
R- 26    0.905085E 01     0.600000E 01
R- 27    0.1C7731E 02     0.100000E 02
R- 28    0.128851E 02     0.160000E 02
R- 29    0.154896E 02     0.180000E 02
R- 3C    0.187166E 02     0.290000E C2
R- 31    0.227358E 02     0.230000E 02
R- 32    0.277657E 02     0.290000E C2
R- 33    0.34C873E 02     0.330000E C2
R- 34    C.420611E 02     0.490000E 02
R- 35    0.521423E 02     0.600000E 02
R- 36    0.648912E 02     0.9400C0E 02
R- 37    0.809695E 02     0.105000E C3
R- 38    0.101C79E 03     0.106000E 03
R- 39    0.125807E 03     0.147000E C3
R- 40    0.155224E 03     0.162000E 03
R- 41    0.188064E 03     0.197000E 03
R- 42    0.220103E 03     0.196000E 03
R- 43    0.241670E 03     0.210000E 03
R- 44    0.235521E 03     0.205000E 03
R- 45    0.181950E 03     0.131000E 03
```

```
R- 46   0.864394E 02    0.7800C0E 02
R- 47   0.134865E 02    0.100000E 02
R- 48   0.100255E 00    0.0
ALPHA ESTIMATES   SIGLEVEL(EACH)=0.050   DFA=    9.   DFE=     27.              LOWER BOUND= 0.514130    UPPER BOUND= 0.925005
NO OF CASES LESS THAN LOWER B= 125    6.25% ; GREATER THAN UPPER B=  88    4.40%  LOWER B(EST)= 0.468784    UPPER B(EST)= 0.922186
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% PLOT OF FREQUENCY DISTRIBUTION %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

CHI-SQ GOODNESS OF FIT TEST: CHI= 0.881404E 02 NDF= 30 PROB= 0.0000



NOTE: (.) EXPECTED FREQ; (*) OBSERVED FREQ; (+) OVERLAPPING POINT

```
@@@@@@@@@@@@@@@@@@@@@@@@@@
@  RELIABILITY STUDY     @
@@@@@@@@@@@@@@@@@@@@@@@@@@
ESTIMATION IS BASED ON KRISTOF CORRECTION(ANOVA,UNBIASED)
EXPECTED FREQUENCY OF RELIABILITY ESTIMATES BELOW -0.2=  0.425398E 01
MAXIMUM=  0.999999E 00   MINUMUM= -0.200000E 00   CLASS INTERVAL=  0.250000E-01
```

DESCRIPTIVE STATISTICS FOR RELIABILITY ESTIMATES

	MEAN OBSERVED	MEAN EXPECTED	VARIANCE OBSERVED	VARIANCE EXPECTED
REL COF	0.813762E 00	0.831249E 00	0.193338E-01	0.143439E-01

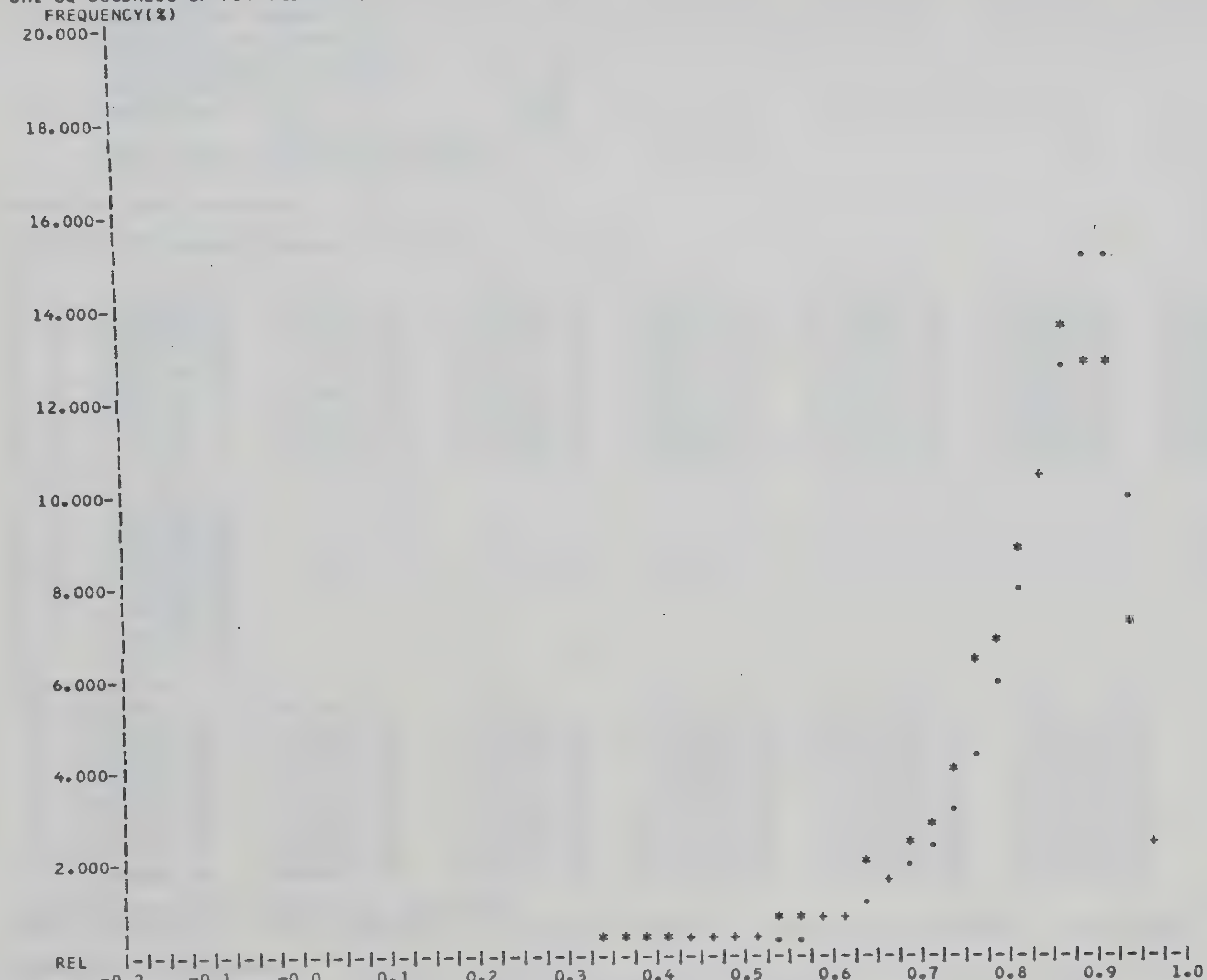
COMPARISON OF RELIABILITY

	C- 1	C- 2
R- 1	-0.252795E 01	0.100000E 01
R- 2	0.154018E 00	0.0
R- 3	0.171065E 00	0.0
R- 4	0.190377E 00	0.0
R- 5	0.212431E 00	0.0
R- 6	0.237584E 00	0.300000E 01
R- 7	0.266075E 00	0.0
R- 8	0.298977E 00	0.0
R- 9	0.336647E 00	0.100000E 01
R- 10	0.380039E 00	0.100000E 01
R- 11	0.430107E 00	0.100000E 01
R- 12	0.488281E 00	0.0
R- 13	0.555754E 00	0.0
R- 14	0.634789E 00	0.200000E 01
R- 15	0.727057E 00	0.0
R- 16	0.835776E 00	0.200000E 01
R- 17	0.963926E 00	0.200000E 01
R- 18	0.111604E 01	0.100000E 01
R- 19	0.129712E 01	0.100000E 01
R- 20	0.151360E 01	0.200000E 01
R- 21	0.177395E 01	0.100000E 01
R- 22	0.208831E 01	0.500000E 01
R- 23	0.247049E 01	0.500000E 01
R- 24	0.293696E 01	0.400000E 01
R- 25	0.351024E 01	0.600000E 01
R- 26	0.421894E 01	0.400000E 01
R- 27	0.510097E 01	0.700000E 01
R- 28	0.620604E 01	0.700000E 01
R- 29	0.760102E 01	0.900000E 01
R- 30	0.937414E 01	0.130000E 02
R- 31	0.116479E 02	0.120000E 02
R- 32	0.145849E 02	0.130000E 02
R- 33	0.184121E 02	0.190000E 02
R- 34	0.234413E 02	0.380000E 02
R- 35	0.301052E 02	0.280000E 02
R- 36	0.390042E 02	0.450000E 02
R- 37	0.509689E 02	0.530000E 02
R- 38	0.671327E 02	0.780000E 02
R- 39	0.889852E 02	0.130000E 03
R- 40	0.118342E 03	0.137000E 03
R- 41	0.156998E 03	0.175000E 03
R- 42	0.205527E 03	0.211000E 03
R- 43	0.259976E 03	0.268000E 03
R- 44	0.304532E 03	0.255000E 03
R- 45	0.301089E 03	0.259000E 03

```
R- 46   0.200904E 03    0.145000E 03
R- 47   0.497977E 02    0.500000E 02
R- 48   0.735044E 00    0.100000E 01
ADJUSTED ALPHA ESTIMATES   SIG LEVEL(EACH)=0.050   DFA=    9.   DFE=      27.      LOWER BOUND= 0.622101    UPPER BOOUND= 0.941671
NO OF CASES LESS THAN LOWER B= 125    6.25% ; GREATER THAN UPPER B=  88    4.40%  LOWER B(EST)= 0.586832    UPPER B(EST)= 0.939478
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% PLOT OF FREQUENCY DISTRIBUTION %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

CHI-SQ GOODNESS OF FIT TEST: CHI= 0.857872E 02 NDF= 25 PROB= 0.0000



NOTE: (.) EXPECTED FREQ; (*) OBSERVED FREQ; (+) OVERLAPPING POINT

*******LAST SEED RANDOM NUMBER IX= 771712087

LORD AND NOVICK PARAMETERS, N=1000, I=15,J=9, LATENT(NO),ERROR(NO)

NO OF SAMPLES SIMULATED	1000
NO OF SUBJECTS IN EACH SAMPLE	15
NO OF ITEMS	9
STARTING SEED RANDOM NUMBER	777
OPTION FOR CARD OUTPUT	1
OPTION FOR PLOT	1
OPTION FOR ESTIMATION FORMULA	2
OPTIION FOR CLASS INTERVALS	48
SIGNIFICANCE LEVEL	0.050

TRUE SCORE DISTRIBUTIONS ARE NORMAL
ERROR SCORE DISTRIBUTIONS ARE NORMAL

FORMAT FOR ITEM PARAMETERS (9F5.5)

INTER ITEM TETRACHORIC CORRELATION MATRIX

	C- 1	C- 2	C- 3	C- 4	C- 5	C- 6	C- 7	C- 8
R- 1	0.240100E 00	0.351330E 00	0.269010E 00	0.290570E 00	0.291550E 00	0.313600E 00	0.233240E 00	0.259700E 00
R- 2	0.351330E 00	0.514089E 00	0.393633E 00	0.425181E 00	0.426615E 00	0.458880E 00	0.341292E 00	0.380010E 00
R- 3	0.269010E 00	0.393633E 00	0.301401E 00	0.325557E 00	0.326655E 00	0.351360E 00	0.261324E 00	0.290970E 00
R- 4	0.290570E 00	0.425181E 00	0.325557E 00	0.351649E 00	0.352835E 00	0.379520E 00	0.282268E 00	0.314290E 00
R- 5	0.291550E 00	0.426615E 00	0.326655E 00	0.352835E 00	0.354025E 00	0.380800E 00	0.283220E 00	0.315350E 00
R- 6	0.313600E 00	0.458880E 00	0.351360E 00	0.379520E 00	0.380800E 00	0.409600E 00	0.304640E 00	0.339200E 00
R- 7	0.233240E 00	0.341292E 00	0.261324E 00	0.282268E 00	0.283220E 00	0.304640E 00	0.226576E 00	0.252280E 00
R- 8	0.259700E 00	0.380010E 00	0.290970E 00	0.314290E 00	0.315350E 00	0.339200E 00	0.252280E 00	0.280900E 00
R- 9	0.242550E 00	0.354915E 00	0.271755E 00	0.293535E 00	0.294525E 00	0.316800E 00	0.235620E 00	0.262350E 00

	C- 9
R- 1	0.242550E 00
R- 2	0.354915E 00
R- 3	0.271755E 00
R- 4	0.293535E 00
R- 5	0.294525E 00
R- 6	0.316800E 00
R- 7	0.235620E 00
R- 8	0.262350E 00
R- 9	0.245025E 00

ITEM PARAMETERS

ITEM	BIS COR	DIFFICULTY	VARIANCE	THRES CONS.	DISC. POWER	DIFF. INDEX
1	0.490000E 00	0.960000E-01	0.867839E-01	0.130488E 01	0.562105E 00	0.266301E 01
2	0.717000E 00	0.199000E 00	0.159399E 00	0.845037E 00	0.102859E 01	0.117857E 01
3	0.549000E 00	0.338000E 00	0.223756E 00	0.417488E 00	0.656838E 00	0.760452E 00
4	0.593000E 00	0.434000E 00	0.245644E 00	0.165862E 00	0.736461E 00	0.279700E 00
5	0.595000E 00	0.471000E 00	0.249159E 00	0.725737E-01	0.740302E 00	0.121973E 00
6	0.640000E 00	0.574000E 00	0.244524E 00	-0.186207E 00	0.832927E 00	-0.290948E 00
7	0.476000E 00	0.676000E 00	0.219024E 00	-0.456114E 00	0.541250E 00	-0.958223E 00
8	0.530000E 00	0.801000E 00	0.159399E 00	-0.845036E 00	0.625002E 00	-0.159441E 01
9	0.495000E 00	0.882000E 00	0.104076E 00	-0.118515E 01	0.569690E 00	-0.239425E 01

POPULATION PARAMETERS UNDER NORMAL OGIVE MODEL
MEAN= 0.447100E 01 VAR= 0.400542E 01 TRUE VAR= 0.265655E 01 ERROR VAR= 0.134888E 01 REL=0.66324 KR20=0.64984

PARALLEL ITEM COVARIANCES

1	2	3	4	5	6	7	8	9
0.84335E-02	0.48553E-01	0.41851E-01	0.55859E-01	0.57335E-01	0.65247E-01	0.30161E-01	0.24188E-01	0.11260E-01

INTER ITEM DISPERSION MATRIX

	C- 1	C- 2	C- 3	C- 4	C- 5	C- 6	C- 7	C- 8
R- 1	0.867839E-01	0.199748E-01	0.178805E-01	0.198965E-01	0.198224E-01	0.198603E-01	0.131971E-01	0.105217E-01
R- 2	0.199748E-01	0.159399E 00	0.434545E-01	0.486104E-01	0.484471E-01	0.485968E-01	0.319959E-01	0.253891E-01
R- 3	0.178805E-01	0.434545E-01	0.223756E 00	0.481019E-01	0.485011E-01	0.504828E-01	0.337193E-01	0.281987E-01
R- 4	0.198965E-01	0.486104E-01	0.481019E-01	0.245644E 00	0.565456E-01	0.595758E-01	0.399012E-01	0.338688E-01
R- 5	0.198224E-01	0.484471E-01	0.485011E-01	0.565456E-01	0.249159E 00	0.607093E-01	0.407533E-01	0.348256E-01
R- 6	0.198603E-01	0.485968E-01	0.504828E-01	0.595758E-01	0.607093E-01	0.244524E 00	0.440617E-01	0.383725E-01
R- 7	0.131971E-01	0.319959E-01	0.337193E-01	0.399012E-01	0.407533E-01	0.440617E-01	0.219024E 00	0.266539E-01
R- 8	0.105217E-01	0.253891E-01	0.281987E-01	0.338688E-01	0.348256E-01	0.383725E-01	0.266539E-01	0.159399E 00
R- 9	0.664337E-02	0.159269E-01	0.181952E-01	0.220045E-01	0.227086E-01	0.252626E-01	0.177800E-01	0.164002E-01

	C- 9
R- 1	0.664337E-02
R- 2	0.159269E-01
R- 3	0.181952E-01
R- 4	0.220045E-01
R- 5	0.227086E-01
R- 6	0.252626E-01
R- 7	0.177800E-01
R- 8	0.164002E-01
R- 9	0.104076E 00

INTER ITEM CORRELATION MATRIX

	C- 1	C- 2	C- 3	C- 4	C- 5	C- 6	C- 7	C- 8
R- 1	0.100000E 01	0.169832E 00	0.128313E 00	0.136271E 00	0.134803E 00	0.136334E 00	0.957223E-01	0.894592E-01
R- 2	0.169832E 00	0.100000E 01	0.230094E 00	0.245659E 00	0.243101E 00	0.246152E 00	0.171240E 00	0.159280E 00
R- 3	0.128313E 00	0.230094E 00	0.100000E 01	0.205174E 00	0.205412E 00	0.215822E 00	0.152316E 00	0.149313E 00
R- 4	0.136271E 00	0.245659E 00	0.205174E 00	0.100000E 01	0.228564E 00	0.243084E 00	0.172023E 00	0.171161E 00
R- 5	0.134803E 00	0.243101E 00	0.205412E 00	0.228564E 00	0.100000E 01	0.245955E 00	0.174453E 00	0.174750E 00
R- 6	0.136334E 00	0.246152E 00	0.215822E 00	0.243084E 00	0.245955E 00	0.100000E 01	0.190395E 00	0.194365E 00
R- 7	0.957223E-01	0.171240E 00	0.152316E 00	0.172023E 00	0.174453E 00	0.190395E 00	0.100000E 01	0.142650E 00
R- 8	0.894592E-01	0.159280E 00	0.149313E 00	0.171161E 00	0.174750E 00	0.194365E 00	0.142650E 00	0.100000E 01
R- 9	0.699026E-01	0.123656E 00	0.119233E 00	0.137620E 00	0.141019E 00	0.158359E 00	0.117764E 00	0.127330E 00

	C- 9
R- 1	0.699026E-01
R- 2	0.123656E 00
R- 3	0.119233E 00
R- 4	0.137620E 00
R- 5	0.141019E 00
R- 6	0.158359E 00
R- 7	0.117764E 00
R- 8	0.127330E 00
R- 9	0.100000E 01

```
@@@@@@@@@@@@@@@@@@@@
@   EXAMPLE RUNS   @
@@@@@@@@@@@@@@@@@@@@
```

DATA MATRIX

	C- 1	C- 2	C- 3	C- 4	C- 5	C- 6	C- 7	C- 8
R- 1	0.0	0.100000E 01	0.0	0.100000E 01	0.100000E 01	0.100000E 01	0.0	0.100000E 01
R- 2	0.0	0.100000E 01	0.100000E 01	0.0	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01
R- 3	0.0	0.0	0.0	0.0	0.0	0.0	0.100000E 01	0.100000E 01
R- 4	0.0	0.0	0.0	0.0	0.0	0.100000E 01	0.100000E 01	0.100000E 01
R- 5	0.0	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01
R- 6	0.0	0.100000E 01	0.100000E 01	0.100000E 01	0.0	0.0	0.0	0.100000E 01
R- 7	0.0	0.0	0.0	0.0	0.0	0.0	0.100000E 01	0.0
R- 8	0.0	0.0	0.0	0.0	0.0	0.100000E 01	0.0	0.0
R- 9	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01
R- 10	0.0	0.0	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.0
R- 11	0.0	0.0	0.0	0.0	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01
R- 12	0.0	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01
R- 13	0.0	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01
R- 14	0.0	0.0	0.0	0.0	0.100000E 01	0.0	0.100000E 01	0.100000E 01
R- 15	0.100000E 01	0.100000E 01	0.100000E 01	0.100000E 01	0.0	0.100000E 01	0.100000E 01	0.100000E 01

	C- 9
R- 1	0.100000E 01
R- 2	0.100000E 01
R- 3	0.100000E 01
R- 4	0.100000E 01
R- 5	0.100000E 01
R- 6	0.100000E 01
R- 7	0.0
R- 8	0.100000E 01
R- 9	0.100000E 01
R- 10	0.100000E 01
R- 11	0.100000E 01
R- 12	0.100000E 01
R- 13	0.100000E 01
R- 14	0.100000E 01
R- 15	0.100000E 01

MSA= 0.663491E 00 MSB= 0.816666E 00 MSE= 0.142064E 00 F= 0.467038E 01 KR20= 0.78588 UNBIASED REL EST(ANOVA)= 0.81647

SAMPLE FIXED EFFECTS VECTOR

1	2	3	4	5	6	7	8	9
-0.48889E 00	-0.88889E-01	-0.88889E-01	-0.88889E-01	-0.22222E-01	0.11111E 00	0.17778E 00	0.17778E 00	0.31111E 00

SAMPLE MEANS VECTOR

1	2	3	4	5	6	7	8	9
0.13333E 00	0.53333E 00	0.53333E 00	0.53333E 00	0.60000E 00	0.73333E 00	0.80000E 00	0.80000E 00	0.93333E 00

GMEAN= 0.622222E 00

SAMPLE DISPERSION MATRIX

	C- 1	C- 2	C- 3	C- 4	C- 5	C- 6	C- 7	C- 8
R- 1	0.123809E 00	0.666667E-01	0.666667E-01	0.666667E-01	-0.142857E-01	0.380953E-01	0.285715E-01	0.285715E-01
R- 2	0.666667E-01	0.266667E 00	0.195238E 00	0.195238E 00	0.857143E-01	0.809524E-01	-0.285714E-01	0.114286E 00
R- 3	0.666667E-01	0.195238E 00	0.266667E 00	0.195238E 00	0.857143E-01	0.809524E-01	0.428572E-01	0.428572E-01
R- 4	0.666667E-01	0.195238E 00	0.195238E 00	0.266667E 00	0.857143E-01	0.809524E-01	-0.285714E-01	0.428572E-01
R- 5	-0.142857E-01	0.857143E-01	0.857143E-01	0.857143E-01	0.257143E 00	0.100000E 00	0.571429E-01	0.571429E-01

R- 6	C.380953E-C1	0.809524E-01	0.809524E-01	0.809524E-01	0.100000E 00	0.209524E 00	0.142858E-01	0.142858E-01
R- 7	0.285715E-01	-0.285714E-01	0.428572E-01	-0.285714E-01	0.571429E-01	0.142858E-01	0.171429E 00	0.285715E-01
R- 8	C.285715E-01	0.114286E 00	0.428572E-01	0.428572E-01	0.571429E-01	0.142858E-01	0.285715E-01	0.171429E 00
R- 9	C.952387E-02	0.380953E-01	0.380953E-01	0.380953E-01	0.428572E-01	0.523810E-01	-0.142857E-01	0.571429E-01

	C- 9
R- 1	0.952387E-02
R- 2	C.380953E-01
R- 3	C.380953E-C1
R- 4	C.380953E-01
R- 5	0.428572E-C1
R- 6	C.523810E-C1
R- 7	-C.142857E-01
R- 8	C.571429E-01
R- 9	0.666667E-01

SAMPLE DISPERSICN : SATURATICN COEFF= 0.36861 HOMOGENEITY COEFF= 0.28968 HOM/SAT= 0.78588

VARIANCE CF ALPHA ESTIMATE UNDER ANOVA= 0.03418 ESTIMATE= 0.01382

VARIANCE CF UNBIASED REL ESTIMATES UNDER ANOVA= 0.02511 ESTIMATE= 0.01015

ESTIMATICN OF POPULATICN PARAMETERS BY PARALLEL METHOD

MEAN	0.445068E 01
VARIANCF	0.398767E 01
TRUE VARIANCE	0.263039E 01
ERROR VARIANCE	0.135728E 01
RELIABILITY	0.659631
KR20	0.646338
NO OF CASES	15015

MEAN VECTOR

1	2	3	4	5	6	7	8	9
0.96037E-01	0.19980E 00	0.33566E 00	0.43157E 00	0.46454E 00	0.57502E 00	0.67239E 00	0.79547E 00	0.88019E 00

PARALLEL ITEM CCVARIANCES

1	2	3	4	5	6	7	8	9
0.87083E-02	0.48487E-C1	0.41387E-01	0.54685E-01	0.55598E-01	0.63736E-01	0.30410E-01	0.26373E-01	0.99993E-02

WITHIN TEST DISPERSION MATRIX

	C- 1	C- 2	C- 3	C- 4	C- 5	C- 6	C- 7	C- 8
R- 1	C.868199E-01	0.203735E-01	0.175817E-01	0.198933E-01	0.189250E-01	0.189699E-01	0.122824E-01	0.991945E-02
R- 2	C.203735E-01	0.159891E 00	0.428272E-01	0.467094E-01	0.491804E-01	0.499484E-01	0.305593E-01	0.260814E-01
R- 3	C.175817E-01	C.428272E-01	0.223009E 00	0.480815E-01	0.493367E-01	0.520097E-01	0.347098E-01	0.282287E-01
R- 4	C.198933E-01	0.467C94E-01	0.480815E-01	0.245333E 00	0.533376E-01	0.629970E-01	0.411540E-01	0.346575E-01
R- 5	C.189250E-01	0.491804E-01	0.493367E-01	0.533376E-01	0.248759E 00	0.592914E-01	0.382344E-01	0.304775E-01
R- 6	C.189699E-01	C.499484E-01	0.520097E-01	0.629970E-01	0.592914E-01	0.244388E 00	0.429300F-01	0.395566E-01
R- 7	C.122824E-01	0.305593E-01	0.347098E-01	0.411540E-01	0.382344E-01	0.429300E-01	0.220295E 00	0.277025E-01
R- 8	C.991945E-02	C.260814E-01	0.282287E-01	0.346575E-01	0.304775E-01	0.395566E-01	0.277025E-01	0.162708E 00
R- 9	C.731126E-02	0.163475E-01	0.183737E-01	0.225386E-01	0.210272E-01	0.236761E-01	0.146291E-01	0.156557E-01

	C- 9
R- 1	C.731126E-02
R- 2	C.163475E-01
R- 3	0.183737E-01
R- 4	C.225386E-C1
R- 5	0.210272E-01
R- 6	C.236761E-01
R- 7	C.146291E-01
R- 8	0.156557E-C1
R- 9	0.105465E 00

```
@@@@@@@@@@@@@@@@@@@@@@@@@
@  SUMMARY OF OUTPUT  @
@@@@@@@@@@@@@@@@@@@@@@@@@
```

MAXIMUM= 0.100000E 02   MINUMUM= 0.0   CLASS INTERVAL= 0.208333E 00

DESCRIPTIVE STATISTICS FOR MEAN SQUARES AND EXPECTED VALUES UNDER ANOVA MODEL

	MEAN		VARIANCE	
	OBSERVED	EXPECTED	OBSERVED	EXPECTED
SUBJECT	0.445152E 00	0.443075E 00	0.193603E-01	0.280450E-01
ITEMS	0.118860E 01	0.118801E 01	0.627478E-01	0.838956E-01
ERROR	0.157094E 00	0.150809E 00	0.261996E-03	0.406131E-03

DISCRIPTIVE STATISTICS FOR FIXED EFFECT ESTIMATES AND EXPECTED VALUES UNDER M.F. MODEL

PART	MEAN	EXPECTED	VARIANCE	EXPECTED
1	-0.398350E 00	-0.398483E 00	0.555131E-02	0.590316E-02
2	-0.294903E 00	-0.294720E 00	0.697102E-02	0.737802E-02
3	-0.158919E 00	-0.158856E 00	0.101086E-01	0.106241E-01
4	-0.629203E-01	-0.629516E-01	0.111262E-01	0.111669E-01
5	-0.299884E-01	-0.299845E-01	0.107425E-01	0.112928E-01
6	0.804681E-01	0.805050E-01	0.102499E-01	0.108360E-01
7	0.177867E 00	0.177874E 00	0.102542E-01	0.109784E-01
8	0.301048E 00	0.300951E 00	0.881981E-02	0.838797E-02
9	0.385696E 00	0.385666E 00	0.692150E-02	0.654620E-02

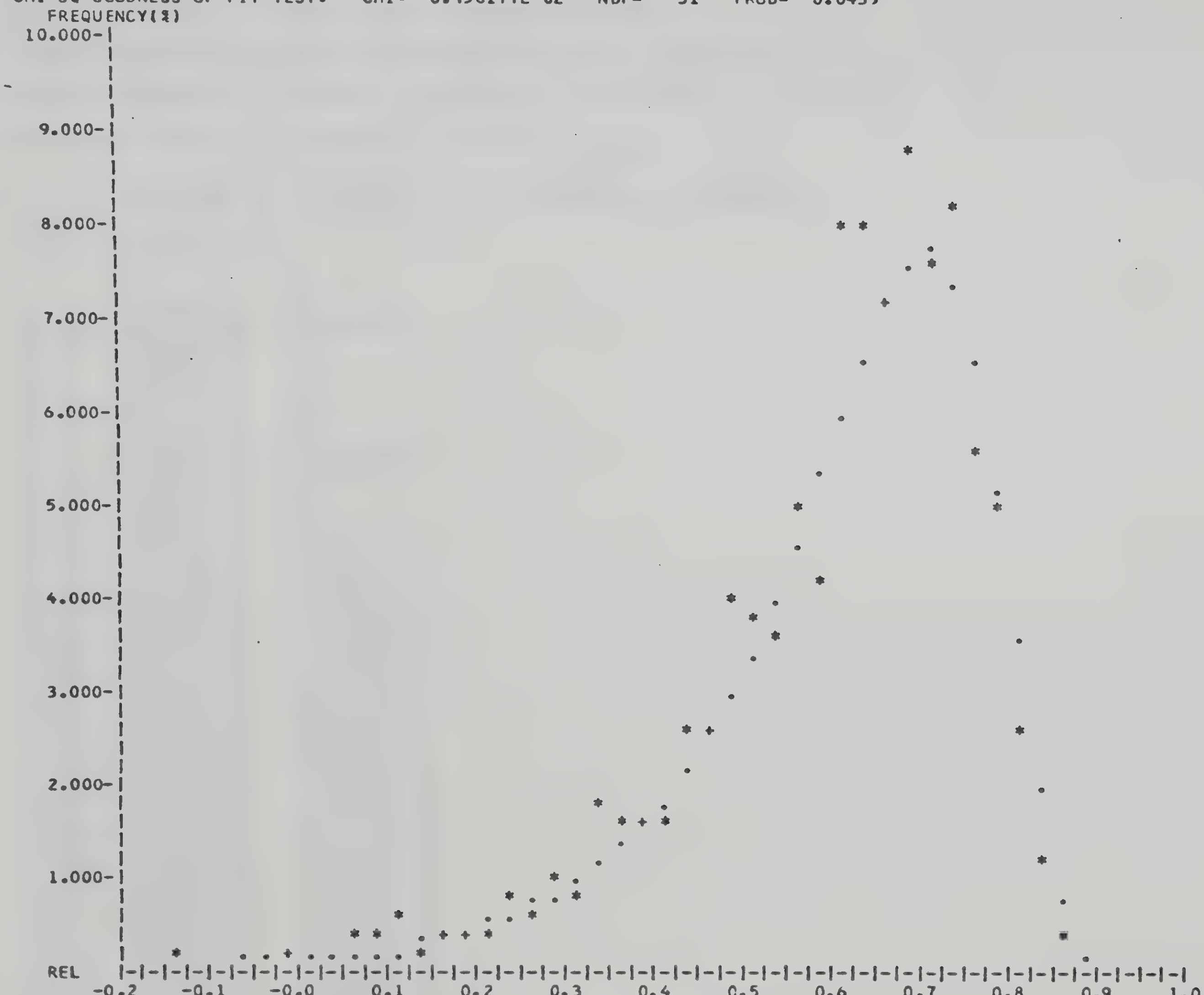
```
@@@@@@@@@@@@@@@@@@@@@@@@@@@
@ RELIABILITY STUDY       @
@@@@@@@@@@@@@@@@@@@@@@@@@@@

ESTIMATICN IS BASED CN ALPHA FORMULA(ANOVA,BIASED)

EXPECTED FREQUENCY OF RELIABILITY ESTIMATES BELCW -0.2=  0.522107E 01

MAXIMUM=  0.999999E 00   MINUMUM= -0.200000E 00   CLASS INTERVAL=  0.250000E-01

DESCRIPTIVE STATISTICS FOR  RELIABILITY ESTIMATES

                     MEAN                   |          VARIANCE
           OBSERVED         EXPECTED        |     OBSERVED        EXPECTED
REL COF    0.599975E 00  0.602903E 00       |     0.319971E-01  0.349162E-01

CCMPARISCN OF RELIABILITY
```

	C- 1	C- 2
R- 1	C.598371E 00	0.0
R- 2	C.676930E 00	0.0
R- 3	C.767112E 00	0.200000E 01
R- 4	C.870764E 00	0.0
R- 5	C.990212E 00	C.0
R- 6	C.112808E 01	0.0
R- 7	C.128729E 01	0.0
R- 8	C.147176E 01	0.100000E 01
R- 9	C.168568E 01	0.0
R- 10	C.193447E 01	0.0
R- 11	0.222385E C1	0.30000CE 01
R- 12	C.256175E 01	C.400000E 01
R- 13	C.295609E 01	0.500000E 01
R- 14	C.341797E 01	0.100000E 01
R- 15	C.395900E 01	0.300000E 01
R- 16	C.459439E 01	C.400000E 01
R- 17	C.534123E 01	0.300000E 01
R- 18	C.621986E 01	0.800000E 01
R- 19	0.725502E 01	0.500000E 01
R- 20	0.847512E 01	C.900000E 01
R- 21	C.991344E 01	0.800000E 01
R- 22	C.116101E 02	0.180000E 02
R- 23	C.136085E 02	0.160000E 02
R- 24	C.159593E 02	0.150000E 02
R- 25	C.187183E 02	0.160000E 02
R- 26	C.219453E 02	0.260000E 02
R- 27	C.257000E 02	0.250000E 02
R- 28	C.300362E 02	C.390000E 02
R- 29	0.349960E 02	0.370000E 02
R- 30	C.405683E 02	0.360000E 02
R- 31	C.467727E 02	C.500000E 02
R- 32	0.534252E 02	0.420000E 02
R- 33	C.602958E 02	0.790000E 02
R- 34	C.669530E 02	C.790000E 02
R- 35	C.727401E 02	0.710000E 02
R- 36	C.767285E 02	0.870000E 02
R- 37	C.777565E 02	0.750000E 02
R- 38	C.745869E 02	0.820000E 02
R- 39	C.662886E 02	0.560000E 02
R- 40	0.528896E 02	C.500000E 02
R- 41	0.361113E 02	0.260000E 02

```
R- 42   C.195656E 02     0.110CCCE 02
R- 43   C.743502E 01     0.300000E 01
R- 44   C.160366E 01     0.0
R- 45   C.133753E 00     C.0
R- 46   0.196695E-02     0.0
R- 47   C.596C46E-04     0.0
R- 48   C.0              0.0
ALPHA ESTIMATES   SIGLEVEL(EACH)=0.050   CFA=   14.   DFE=    112.              LOWER BOUND= 0.257551    UPPER BOUND= 0.808942
NO OF CASES LESS THAN LOWER B=  40    4.00% ; GREATER THAN UPPER B=  29    2.90% LOWER B(EST)= 0.290827   UPPER B(EST)= 0.796666
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% PLOT OF FREQUENCY DISTRIBUTION %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

CHI-SQ GOODNESS OF FIT TEST: CHI= 0.456177E 02 NDF= 31 PROB= 0.0439



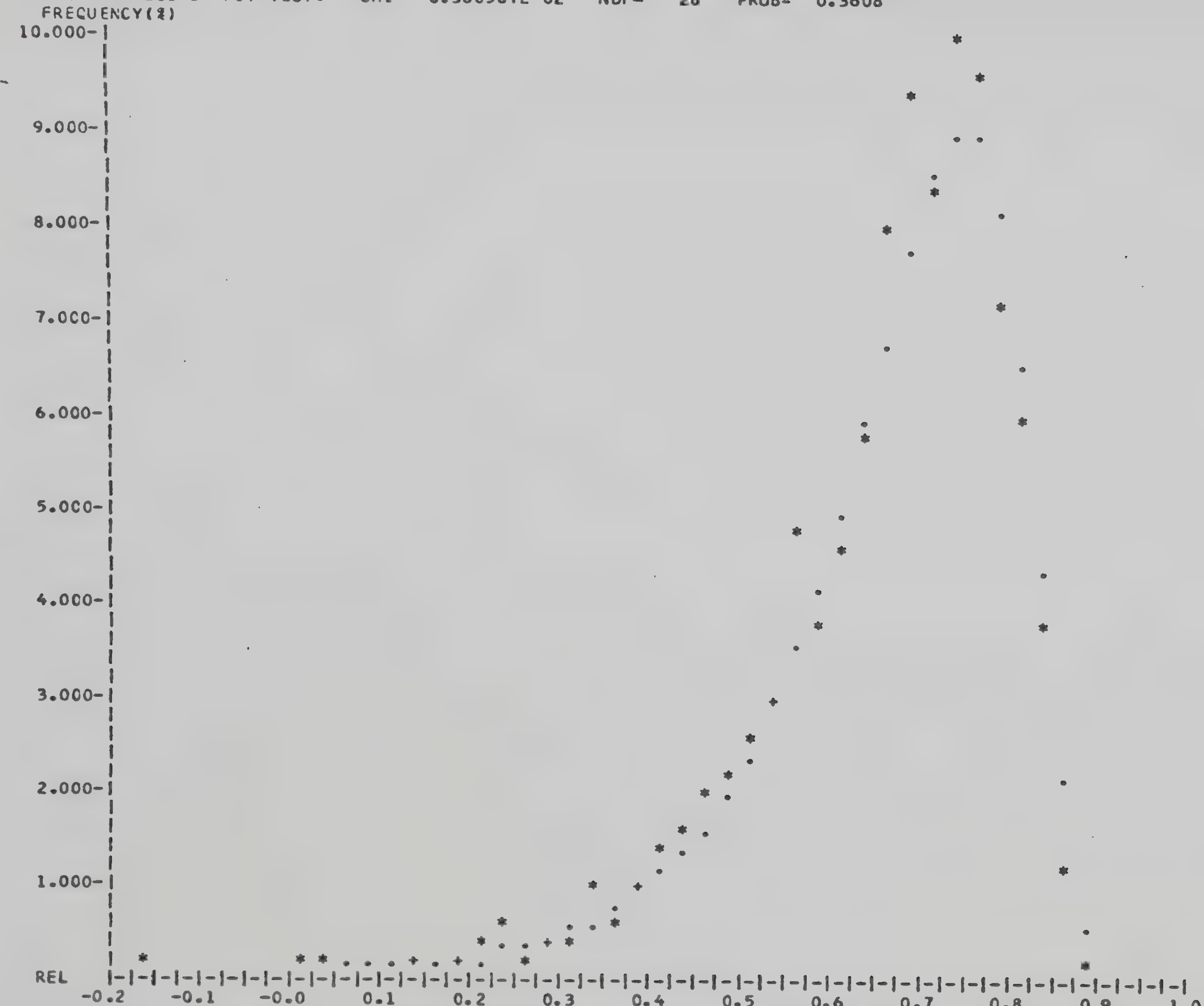
NOTE: (.) EXPECTED FREQ; (*) OBSERVED FREQ; (+) OVERLAPPING POINT

```
@@@@@@@@@@@@@@@@@@@@@@@@@@
@  RELIABILITY STUDY     @
@@@@@@@@@@@@@@@@@@@@@@@@@@

ESTIMATICN IS BASFD CN KRISTOF CORRECTICN(ANOVA,UNBIASED)

EXPECTED FREQUENCY OF RELIABILITY ESTIMATES BELCW -0.2=  0.522107E 01

MAXIMUM=  0.999999E 00   MINUMUM= -0.200000E 00   CLASS INTERVAL=  0.250000E-01

DESCRIPTIVE STATISTICS FOR  RELIABILITY ESTIMATES

                     MEAN                  |            VARIANCE
            OBSERVED        EXPECTED       |     OBSERVED        EXPECTED
REL COF     0.657105E 00  0.659631E 00     |     0.235255E-01  0.256527E-01

CCMPARISCN OF RELIABILITY

          C-  1          C-  2
R-  1  -C.263560E 01    0.0
R-  2   C.316441E 00    0.100000E 01
R-  3   0.361085E 00    0.0
R-  4   C.412703E 00    0.0
R-  5   C.472844E 00    0.0
R-  6   C.542819E 00    0.0
R-  7   C.624478E 00    0.0
R-  8   C.719965E 00    0.0
R-  9   0.832021E 00    0.100000E 01
R- 10   C.963688E 00    0.100000E 01
R- 11   0.111884E 01    C.0
R- 12   C.130212E 01    0.0
R- 13   C.151902E 01    0.0
R- 14   C.177658E 01    0.100000E 01
R- 15   0.208282E 01    0.0
R- 16   C.244838E 01    C.200000E 01
R- 17   C.288522E 01    0.400000E 01
R- 18   C.340885E 01    C.600000E 01
R- 19   0.403786E 01    0.100000E 01
R- 20   0.479501E 01    0.400000E 01
R- 21   C.570798E 01    C.300000E 01
R- 22   C.681144E 01    C.900000E 01
R- 23   C.814652E 01    C.600000E 01
R- 24   C.976372E 01    0.100000E 02
R- 25   0.117233E 02    C.130000E 02
R- 26   0.140983E 02    0.150000E 02
R- 27   C.169718E 02    0.200000E 02
R- 28   C.204423E 02    C.210000E 02
R- 29   0.246159E 02    0.250000E 02
R- 30   C.296031E 02    0.300000E 02
R- 31   0.355071E 02    0.480000E 02
R- 32   0.424030E 02    0.380000E 02
R- 33   0.502901E 02    0.460000E 02
R- 34   C.590509E 02    0.570000E 02
R- 35   0.683381E 02    0.790000E 02
R- 36   C.774779E 02    C.940000E 02
R- 37   C.853160E 02    0.840000E 02
R- 38   C.901400E 02    C.990000E 02
R- 39   C.897648E 02    0.950000E 02
R- 40   C.820366E 02    0.720000E 02
R- 41   C.660273E 02    0.590000E 02
```

```
R- 42   0.438172E 02    0.380000E 02
R- 43   C.214796E C2    0.120000E 02
R- 44   C.641847E 01    0.200000E 01
R- 45   C.820398E 00    0.0
R- 46   0.218749E-01    0.0
R- 47   C.0             0.0
R- 48   C.596046E-04    0.0
ADJUSTED ALPHA ESTIMATES   SIG LEVEL(EACH)=0.050   DFA=   14.   DFE=    112.     LOWER BOUND= 0.363615    UPPER BOOUND= 0.836236
NO OF CASES LESS THAN LOWER B=  40    4.00% ; GREATER THAN UPPER B=  29    2.90%  LOWER B(EST)= 0.392137   UPPER B(EST)= 0.825714
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% PLOT CF FREQUENCY DISTRIBUTION %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

CHI-SQ GCCDNESS OF FIT TEST: CHI= 0.300507E 02 NDF= 28 PROB= 0.3608



NOTE: (.) EXPECTED FREQ; (*) CBSERVED FREQ; (+) OVERLAPPING POINT